L.E. GUREVICH

THE MAGIC OF GALAXIES AND STARS

L.E. GUREVICH
A.D. CHERNIN

THE MAGIC OF GALAXIES AND STARS

Preface

This book covers the cosmogony of stars and galaxies, a new field in astrophysics. The modern development of cosmogony is related to the astronomic discoveries of the past two decades, starting from the discovery of quasars in 1963 and microwave background radiation in 1965. The researchers in this area proceed from the achievements of cosmology, a science dealing with the universe as a whole, and make use of data from different branches of astronomy, physics, and mathematics.

The problems of cosmogony are diverse and complicated, and not every one of them has been solved yet. However, there is a number of reliable observational data and theoretical conclusions related to the formation of stars and star systems, and the general pattern of the origination of the large-scale structure of the universe is only gradually becoming clearer.

This book deals with the latest achievements of cosmogony, its problems and prospects, using the simple language of school physics and astronomy. In fact, the basic ideas and hypotheses allow a demonstrative presentation without mathematical formulas, and we hope that the general reader, keen on science news, will be eagerly interested.

This popular-science book follows our monograph *Introduction to Cosmogony* (Nauka, Moscow, 1978, in Russian), which reviewed and analyzed the investigations of the authors and their colleagues and gave a general presentation of the modern science of cosmogony. The principal ideas of the monograph are reflected in this book as well; furthermore, we added new material on recent star formation, the final stages of their evolution, and the role of neutrinos in cosmogony.

We are deeply grateful to our colleagues V. A. Antonov, A. S. Zentsova, A. S. Zilbergleit, V. A. Ruban, E. A. Tropp, and A. Yu. Ushakov. A number of points and ideas presented in this book have been developed in our joint research with them.

We extend our sincere gratitude to A. S. Zilbergleit, I. D. Novikov, E. A. Tropp, and A. V. Tutukov, who read the manuscript and offered their helpful comments.

L. E. Gurevich, A. D. Chernin

Chapter 1

The Universe

The age of the Sun is more than 5000 million years, the oldest stars in the universe are over 10,000 million years, and the youngest stars are being born and beginning their evolution right now in the clouds of gas and cosmic dust. Stars are incorporated into systems of various masses and dimensions: star pairs, groups, complexes, associations, clusters, and galaxies. Galaxies are not at the top of the hierarchy of astronomical systems, for there are also galactic clusters and superclusters, the largest formations known so far in the universe.

The history of cosmic structures spans 12,000-15,000 million years, and the age of the universe is no less than 15,000-18,000 million years. Prior to the formation of the present-day planets, stars, and galaxies, their matter was just hot hydrogen-helium plasma uniformly distributed throughout the universe. Many centuries of studies on the structure and evolution of celestial bodies, observational discoveries of the 20th century, and especially the discovery of the expansion of the universe and the microwave background radiation in it suggest certain ideas on the properties of the space medium in the prestar, pregalactic epoch, on the physical processes resulting in the formation of the observed structure in the universe, and on the continuing cosmogonical process.

The first chapter deals with the general outlines of the modern astronomical understanding of the universe and the principal ideas of cosmology, i.e. the science of the universe as a whole.

Stars and Galaxies

The Sun is one of the 100,000 million stars incorporated into our Galaxy, a gigantic star system which we can see in the sky as the white band of the Milky Way (Fig. 1). The Galaxy consists of a flat subsystem, which looks like a disk with a bulge in the middle, and the spherical subsystem, within which this disk is located (Fig. 2). The disk and the spherical subsystem of the Galaxy contain approximately the same number of stars. The Sun belongs to the galactic disk, and the distance from the Sun to the galactic centre is two thirds of the disk's radius. The radii of the disk and the spherical subsystem are almost the same, 15 kiloparsecs (1 parsec (pc) is about three light years, or 3×10^{18} cm, 1 kiloparsec (kpc)=1000 pc).

Fig. 1
A region in the Milky Way containing stars and gas-dust clouds (the California Nebula).

In addition to stars, the galactic disk contains interstellar gas and cosmic dust whose mass is only several per cent of that of the stars; the spherical subsystem does not actually contain any gas and dust. There is a noticeable number of young bright stars among those of the disk, while such stars are almost absent in the spherical subsystem. The galactic disk rotates on the whole, although the angular velocity is different at different distances from the disk's centre. The linear velocity of the disk's rotation is about 220-250 km/s in the vicinity of the Sun. The stars of the disk revolve around its centre along almost circular orbits, and the deviations from this circular motion do not exceed 20 km/s. The velocity of the general rotation of the spherical subsystem's stars (around the centre

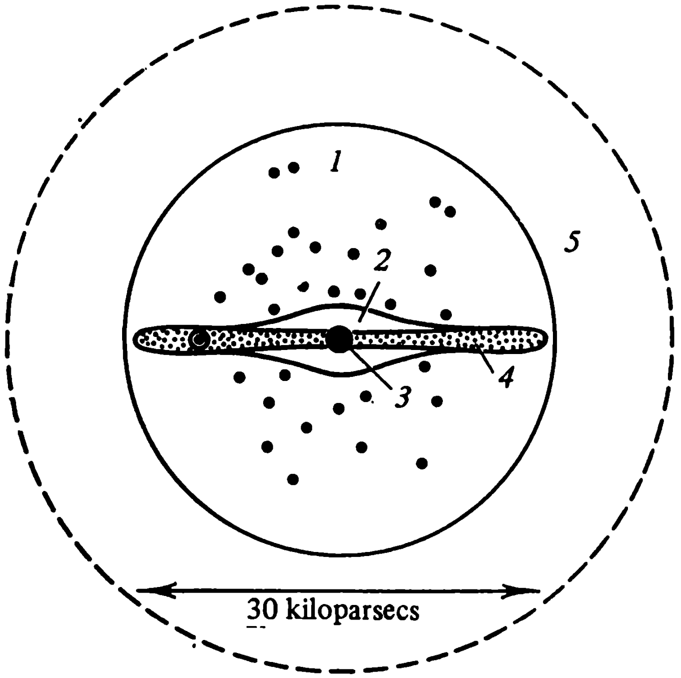

Fig. 2
A diagram of the Galaxy's structure. The points represent some of the globular clusters. The position of the Sun is marked with the sign ⊙, which can be found in ancient Egyptian inscriptions. *1*—the spherical subsystem; *2*—the disk; *3*—the nucleus; *4*—the layer of gas-dust clouds; *5*—the corona. The dimensions are conditional. The corona's radius is really several times greater than the disk's radius.

of the Galaxy) in the vicinity of the Sun is at least five times less than that of the disk's stars. The stars of the spherical subsystem follow elongated orbits, and their typical velocities are 2-3 hundred km/s.

A considerable part of the disk's stars is incorporated into various groups. No less than half of all stars are binaries, i.e. star pairs, and large formations are distant clusters containing up to a thousand stars attracted to each other by mutual gravitation. The youngest stars of the disk and the clouds of gas and dust are located in the spiral arms, i.e. the wide and bright bands originating in the central area of the Galaxy.

The distribution of stars in the spherical subsystem is more or less spherically symmetrical. About a thousandth part of them belong to large clusters containing up to a million stars and are called globular clusters (Fig. 3).

Within both subsystems of the Galaxy, the number of stars per unit space increases towards the central area, i.e. the Galaxy's nucleus, which is a source of intense radio, infrared, X-ray, and gamma-ray emanation. The nucleus also appears to be a source of gas emission.

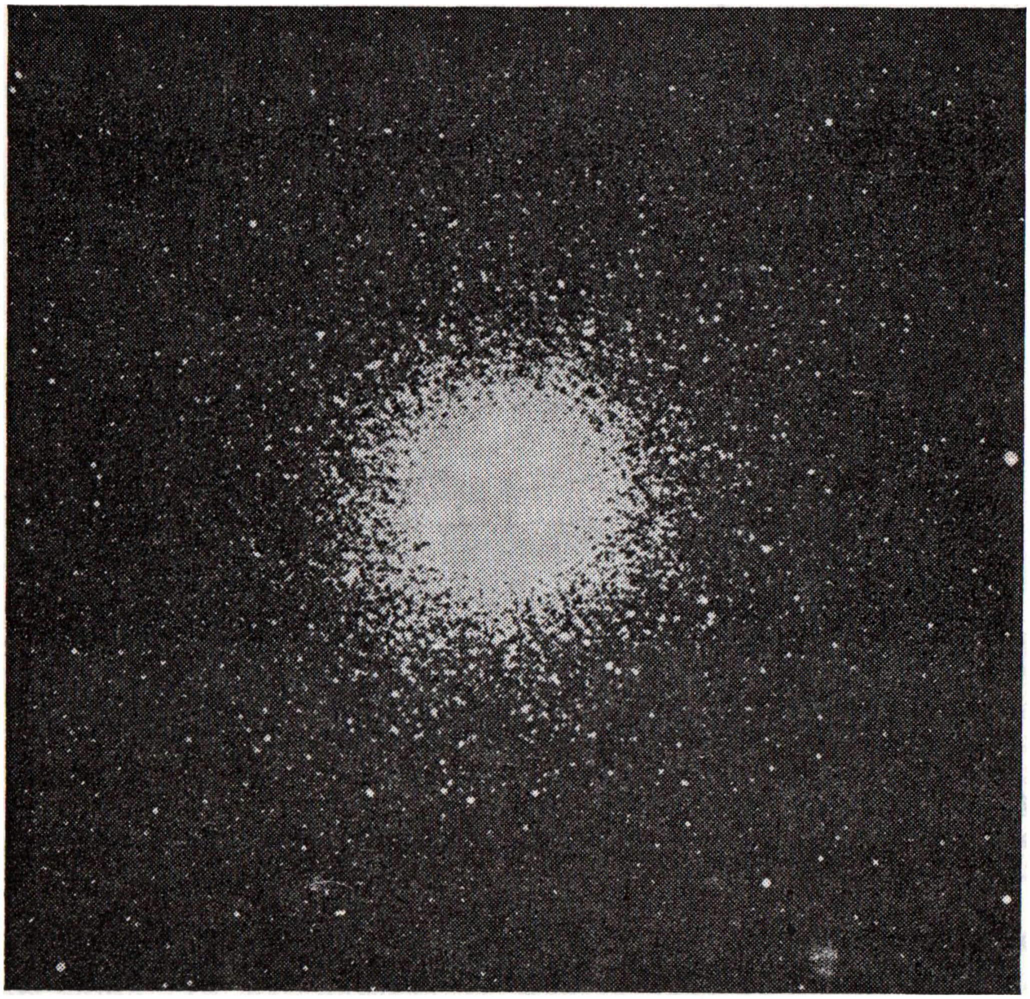

Fig. 3
A globula cluster in the Hercules constellation.

The luminosity of the Galaxy, i.e. the total energy radiated by all its stars per unit time, is 3×10^{37} W. This is about 100,000 million times that of the Sun's luminosity (4×10^{26} W). The total mass of the Galaxy's stars is estimated at 2×10^{44} g, which amounts to 100,000 million Sun masses (2×10^{33} g). Both astronomy and astrophysics widely employ the Sun mass and luminosity as a measure of the masses and luminosities of stars and star systems.

It has recently been found that the Galaxy is surrounded by an extended corona stretching dozens of times farther from the centre than the disk and the spherical subsystem. The total mass of the corona is several times that of all the Galaxy's stars put together, but because of large dimensions the corona's density is much smaller than that created by stars and gas-dust clouds. The corona makes itself felt by its gravitation, but it does not emit any light, and there are neither stars nor clouds in it. The coronas surrounding galaxies and their "hidden masses" will be discussed in more detail in Chapter 6.

There is a great number of other star systems in the universe, i.e. galaxies similar to our Galaxy. The galaxies possessing a disk subsystem with a spiral pattern are called spiral. The famous Andromeda Nebula (see Fig. 17) is an immense spiral galaxy nearest to us. Its mass and luminosity are each two times those of our Galaxy. Other spiral galaxies are not so massive; most commonly their masses are 1000-10,000 million Sun masses, while their luminosities are 10-100 times less than that of our Galaxy.

Besides spiral galaxies, there are elliptical galaxies, whose structure and star population are similar to the spherical subsystem of our Galaxy. There is practically no gas-dust matter or young bright stars in them. The largest elliptical galaxies possess masses and luminosities ten times those of our Galaxy. There are also dwarf elliptical galaxies with masses and luminosities ten thousand times less (see Fig. 46). An elliptical galaxy, in particular a very massive one, very often possesses a dense nucleus which is usually larger and more active than that of a spiral galaxy.

There are also typically irregular galaxies. Their masses and luminosities are ten times less than those of our Galaxy. Their star population is similar to that of the disks of spiral galaxies. However, the stars and considerable masses of gas-dust matter in them do not define a regular structure and do not possess a pronounced general rotation. Besides bright young stars, irregular galaxies contain old and less bright stars (similar to those of the Galaxy's spherical subsystem), which also define a basically spherical structure.

These three types of galaxies were first identified and investigated by E. Hubble and other astronomers during the 1920's-1930's. Galaxies of other types, not always falling into the original classification, have become known since that time. They include primarily the galaxies with active nuclei and considerable radio emission. The most extraordinary objects of this type are quasars, i.e. quasi-stellar radio sources, discovered during the 1960's. Stars cannot be detected there; they are either absent, or, which is more probable, they cannot be discerned against the colossal luminosity of a quasar's nucleus, which reaches 10^{39}-10^{40} W, i.e. tens of thousands of times greater than that of our Galaxy. This energy is radiated from areas 10^{16}-10^{18} cm in size, which is tens and hundreds of thousands of times less than the dimensions of our Galaxy. The radio emission of a quasar is comparable in its intensity to the optical emission, while its infrared radiation is often greater. There is a variety of quasars with low radio emission; such objects are called quasags, i.e. quasi-galaxies.

Owing to their immense luminosity, quasars can be observed from very great distances. The most remote objects which can be viewed through modern astronomical instruments are naturally quasars. They define, as it were, the boundaries of the Metagalaxy, i.e. the observed area of the universe. The distance to the most remote quasars is estimated at thousands of megaparsecs (1 megaparsec (Mpc)=1,000,000 pc). The light from them travels to us for thousands of millions of years.

Most galaxies are in groups or clusters of tens to thousands. There are clusters of galaxies of regularly spherical or ellipsoidal shape; for instance, one of the greatest known clusters, which is in the Berenice's Hair (Coma) constellation and has a radius of about 4 Mpc, contains about ten thousand galaxies, most of them elliptical.

As it has been discovered during recent years, many rich clusters of galaxies contain considerable quantities of hot gas, which reveals itself in its X-ray radiation. The gas temperature reaches a hundred million kelvins, so the gas is in the state of plasma, i.e. it is ionized to such a degree that electrons are separated from nuclei. The mass of hot gas in clusters is comparable to the total mass of galaxies themselves. Judging by the dynamics of galaxies in clusters and the temperatures of the intergalactic gas, these systems contain three to ten times greater quantities of other matter, which only reveals itself by the gravitation it produces. These "hidden masses", which have been mentioned above in connection with galactic coronas, will be discussed in Chapter 6.

Groups and clusters of galaxies are distributed not quite randomly in the universe. The Local Group of galaxies includes our Galaxy, the Andromeda Galaxy, and about 30 lesser objects. The Local Group and two or three other close groups of galaxies define a system called the Local Supercluster. This is a flat formation up to 50 Mpc in size; its plane is perpendicular to that of the disk of our Galaxy. The centre of the Local Supercluster is in the direction of the Virgin (Virgo) constellation, in a major cluster of galaxies at a distance of 20 Mpc from us. Other superclusters are known to be 20 to 100 Mpc in size; their masses are 10^{15} to 10^{16} Sun masses.

Looking at a large-scale map of the sky where galaxies are just points, we can often see clusters of galaxies as extended chains, probably superclusters. The chains are connected and cross each other, producing a cell or honeycomb structure. Whether the cell "superstructure" is universal is yet to be verified by observations, but several cells have already been studied reliably.

The hierarchy of space structures is topped by clusters and superclusters. No larger formations are found in the Metagalaxy.

Counting the number of galaxies in great volumes of 300 Mpc and more in size, which contain many clusters and superclusters, we can find their average concentration in space; knowing galactic masses, we can also estimate the average density of matter in such volumes. This density proves to be the same wherever we choose to take such a volume in space. According to present-day data, it is 3×10^{-31} g/cm^3 or, in terms of hydrogen atoms, about one atom per 30 m^3.

However, astronomical mass estimates are not very reliable. The problem is complicated by the fact that besides the luminous matter in the galaxies there appear to exist considerable masses of matter in the space around them. We cannot observe these masses directly; possibly, they are gases or stars of low luminosity or even black holes; they may also be neutrinos (if they possess a rest mass, see Chapter 6). As we mentioned, hidden masses only reveal themselves in their gravitation which influences the motion of galaxies in their groups and clusters. This gives ground to estimate the related average density; in the opinion of. J. E. Einasto and his colleagues of the Tartu observatory (USSR), it may be two to three times or even five to ten times greater than the average density of the observed galaxies.

The fact that the number of galaxies and the density of matter turn out to be uniform over sufficiently great volumes, wherever these volumes may be taken, implies that the universe is homogeneous on the average if it is considered on a large-scale basis. This is one of the fundamental properties of the world around us.

Cosmological Expansion

Another fundamental property of the universe is its general expansion. Observations show that clusters (and superclusters) of galaxies, being at distances of 100-300 Mpc, move away from each other. This fact was established by E. Hubble in the late 1920's.

It has long since been known that when a source of sound moves away from us, the frequency of the sound we hear decreases, and conversely, when the source moves towards us, the frequency increases. A similar phenomenon occurs in the propagation of light or any other electromagnetic waves. This phenomenon is called the Doppler effect. When a source of light moves away from an observer, the frequency of the observed light also decreases. The colour changes, for instance, from blue to yellow or from yellow to red. Registering the light from distant galaxies, E. Hubble established that the spectral lines in their radiation are shifted to the red part of the spectrum. The farther the galaxy, the greater the "red shift" of the arriving radiation. It follows that galaxies move away from us, and the velocity of recession is greater, the farther away a galaxy is. However, our own Galaxy, where we carry out our observations, is not the centre of the universe, and we have apparently to assume that galaxies or, more accurately, clusters of galaxies generally recede from each other rather than withdraw from us.

If the distance between clusters is L, the velocity of their mutual recession $v = HL$. This relationship is called the Hubble law (or the law of redshift); H is the Hubble constant, and its value does not depend on the position of clusters in space. According to present-day estimates, $H = 55\text{-}75$ km/(s·Mpc).

The instability of the universe was predicted by A. A. Friedmann, the founder of modern cosmology, several years before Hubble's discovery. Proceeding from Einstein's general theory of relativity, Friedmann developed a model of a uniform universe, which proved to be unable to remain at rest and must be unstable. This instability reveals itself in the recession of galaxies and their clusters. It occurs in such a manner that the general uniformity in the distribution of clusters (or superclusters) holds true. As theory shows, the preservation of uniformity demands that the velocities of the recession of bodies from each other be proportional to the distances between them; this is precisely what is found in astronomical observations.

The velocities of the cosmological expansion are considerable. If a cluster of galaxies is at a distance of, for instance, a thousand megaparsecs from us, then, according to the Hubble law, it moves away from us with a velocity no less than 55,000 km/s. The farthest quasars recede with velocities scarcely less than the velocity of light, which is equal to 300,000 km/s.

The expansion occurs at great velocities, while universal gravitation, i.e. the mutual attraction between cosmic systems, struggles to turn the expansion into contraction. The gravitation is greater if the masses of systems are greater and the distances between them are less, and therefore the expansion depends on the density of matter in the universe. This density should be sufficiently large, i.e. exceed a certain critical value (Fig. 4), for the gravitation to overcome the expansion.

The critical density can be found by estimating the energy of recession by the observed velocities of the cosmic systems. Modern data give values from 10^{-29} to 5×10^{-30} g/cm^3, which corresponds to about ten or five hydrogen atoms per cubic metre. This is greater than the average density of galaxies, but does not seem to exceed the density which could be contributed by "hidden masses".

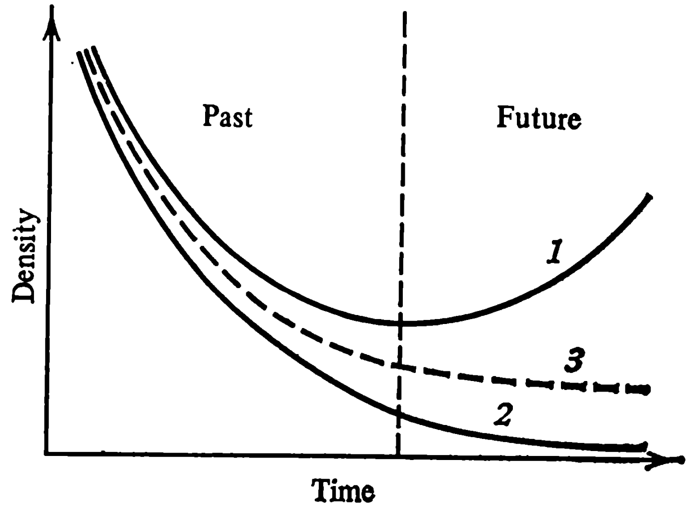

Fig. 4
Three models of the cosmological expansion. *1*—the density of the universe is greater than the critical density; *2*—the density of the universe is less than the critical density; *3*—the density of the universe equals the critical density.

Consequently, the future of the cosmological expansion remains questionable: if the "hidden masses" really exist and their density conforms with their maximum estimate, then the initial "acceleration" was insufficient for the expansion of the universe to continue indefinitely. Then gravitation will be able to put an end to the expansion and bring the universe to contraction in 10,000-15,000 million years. We shall return to this problem in Chapter 6.

Now let us discuss the universe's past, its history rather than its future. While cosmic systems recede from each other now, they were closer in the past or even "contacted" each other. During still earlier times, neither clusters nor galaxies nor even stars could evidently have existed in their present-day condition, and the matter which comprises them should have been uniformly mixed in a single cosmic medium. The farther into the past we go, the greater the density of the medium.

If we could gaze into the past, when would this increase in the density stop? According to Friedmann, the density of the universe increases without any limit in the past and becomes however great, or infinite, at a certain moment. This moment is taken in Friedmann's theory as the point of origin, or zero time. Anything "earlier" than this instant is beyond the Friedmann model, and the model cannot be applied to the very instant of zero time and infinite density either.

The history of physics reiterates that when an infinity appears in theoretical models or formulas, this implies that there is a novel phenomenon fundamentally different from the one that the very models and formulas describe. For instance, an infinity appeared in aerodynamic formulas when the velocity of a body approached the velocity of sound in the medium where the body moved; the resistance of the medium to such motion turned out to be infinite. This would mean that supersonic motion would be impossible. But we well know that aeroplanes can fly with velocities exceeding the velocity of sound in air. The point is that the aerodynamic formulas mentioned above described the resistance in a continuous medium, without abrupt jumps in density and pressure. However, the transition from subsonic to supersonic motion is associated with violating this condition: a shock wave appears in the medium in front of the body, and there occurs a jump in the density and pressure of the medium at the front of the wave. This phenomenon was taken into account. Aerodynamics was reviewed to include the case of the discontinuity of the medium, and the infinity disappeared from theoretical formulas: they gave correct and finite values for the resistance to supersonic motion.

There is no doubt that the infinite density in cosmology implies that something specific should occur at the zero instant. It should have been a colossal phenomenon (the "Big Bang") giving enormous velocities of recession to all matter in the universe. The nature of this phenomenon, i.e. the cause of the cosmological expansion, remains yet unknown.

However, what occurred after the beginning of the cosmological expansion (at a finite although extremely great density), i.e. the dynamics of the cosmological expansion, the physical processes in the expanding matter, can be reliably investigated on the basis of general laws of physics and the Friedmann model, which is in keeping with them.

Friedmann's cosmology reveals the behaviour of the universe's density in time. It turns out that the age of the universe, i.e. the time t from the beginning of the expansion to a certain moment in the history of the universe, is related to the density ρ at this instant by an approximate relationship $t \approx 1/\sqrt{G\rho}$ (here $G = 6.7 \times 10^{-8}$ cm^3/(g$\cdot$s^2) $= 7 \times 10^{-11}$ N$\cdot$m^2/kg^2 is the gravitational constant). Refined formulas are $t = 1/\sqrt{32\pi G\rho/3}$ at t less than one million years and $t = 1/\sqrt{6\pi G\rho}$ at t greater than one million years. The universe's density at the age of one second amounted to approximately 5×10^5 g/cm^3; the density comparable to that of an atomic nuclei, i.e. 10^{15} g/cm^3, occurred at the age of 20 microseconds, and the density of 1 g/cm^3 was reached 10 minutes after the beginning of the expansion.

The relationship between time and density was obtained in the Friedmann model as a result of solving complicated equations of the general theory of relativity. But when it became known, it was soon evident that this simple relationship could be obtained proceeding from an analysis of dimensions in the theory of physical values. This is a powerful method in theoretical physics, and it proceeds from the simple fact that if the left-hand side of an equation contains a value, for instance of time or mass, then the values on the right-hand side of the equation should possess the same dimension, i.e. they should also be values of time or mass. When a theory operates with a small number of dimensional values, it sometimes becomes possible to find relationships between them simply by demanding the same dimensions on the left- and right-hand sides of equations. Here we have the relationship between time and the density of the gravitating medium, no forces other than gravitation being present. It is natural to expect that there is such a relationship between time and density into which, besides these two values, only the value of the gravitational constant enters, because it reflects the presence of the force of gravitation. It is easy to see that these three values can produce one and only one combination satisfying the consideration of dimensions, and this is the relationship of the Friedmann cosmological model. We shall deal with this relationship repeatedly.

The Geometry of the Universe

The properties of space and time cannot be given once and for all, they are not absolute but depend on the distribution and motion of gravitating masses. This is the central idea of Einstein's general theory of relativity and relativistic physics in general. Therefore space, unifoımly filled with matter, should be uniform itself, i.e. everywhere identical in its geometrical properties. If the property of uniformity, which has been established in observations of a part of the universe, encompasses the entire universe, we can make the reasonable assumption that there is nothing special about the observed part, and the universe is the same everywhere.

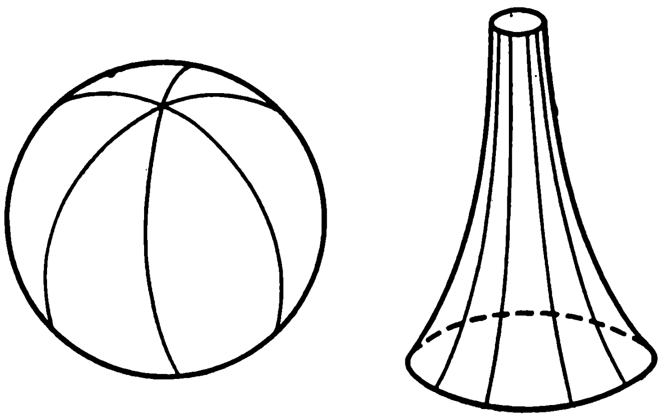

Fig. 5
A sphere and a pseudosphere as examples of curved surfaces with non-Euclidean geometry. The surface of a sphere is finite, and that of a pseudosphere is infinite (the figure shows only a part of a pseudosphere).

Physical space is uniform, but it can be curved by the gravitation of matter. Three-dimensional space can be curved and become non-Euclidean, like two-dimensional surfaces, e.g. the surface of a sphere or a pseudosphere (Fig. 5), can be curved. Although any graphic representation of curved three-dimensional spaces is difficult, a transition to them from curved two-dimensional surfaces can be imagined by proceeding from the transition from a plane (a Euclidean two-dimensional surface) to an ordinary Euclidean three-dimensional space, and by transferring the properties of two-dimensional prototypes to their three-dimensional counterparts. Thus, we can expect that a three-dimensional analogue of a sphere should have a finite volume just like a sphere has a finite surface area. This is an example of a finite three-dimensional space. However, a three-dimensional analogue of a pseudosphere should possess an infinite volume since the total area of a pseudosphere is infinite. This is an example of an infinite space. An intermediate case is that of a noncurved Euclidean three-dimensional space. It also has an infinite volume since the area of an unlimited plane is infinite.

Then what is the real space of the homogeneous universe? An answer could be provided by direct geometrical measurements with the aid of the rays of light from galaxies. However, the deviations from the properties of the usual Euclidean space cannot yet be traced from the distances at which the galaxies are observed. Space could only exhibit its curvature at far greater distances. But such measurements are infeasible: there are few bright sources of light there.

There is another solution to the problem proceeding from the general concepts of relativistic physics concerning the relationships between the geometry of space and gravitating masses in it. If the mutual recession of cosmic bodies, i.e. clusters and superclusters of galaxies, is going to last without any limitation, and the distances between the bodies are going to increase infinitely over the course of time, then, according to Friedmann's theory, the volume of space where they move should also be infinite. The geometrical properties of such a space should be similar to those of a pseudosphere. Conversely, limited expansion in time and the possibility of contraction in the future imply that the universe, which is in this case similar to a sphere, is finite. The reader can find more details on the geometry of the universe in the books by L. E. Gurevich and A. D. Chernin (1970) and I. D. Novikov (1983) listed in Recommended Literature.

The Horizon

When it is said that matter is uniformly distributed in space, it is implied that the overall picture of the recessing cosmic systems is observed as if simultaneously.

Indeed, since the density of matter in the universe decreases because of the cosmological expansion, we can only assume the average density of the universe to be the same everywhere provided that each part of the universe is considered at the same stage of expansion: this is what "simultaneous" means in this case. Otherwise, a part would look denser, i.e. "younger", and another part less dense, i.e. "older", if we saw the first part at an earlier stage of expansion and the second part at a later stage. It can be said that the uniformity of density can only be revealed in a "snapshot" of the universe taken in such imaginary "rays" that propagate instantaneously, with an infinite velocity, and this would be precisely the picture in which the universe on the whole would be uniform.

Naturally, there are no instantaneously propagating waves or rays; any signal propagates with a finite velocity, and the greatest velocity is that of light. What is the universe like if photographed in real light rays?

Astronomers perform their observations in optical, radio, infrared, ultraviolet, X-ray, and gamma-ray ranges of electromagnetic waves. The waves show the pattern of the sky with a lag because they take a certain time to cover the distance between an observed object and an observer. We see the Sun with a lag of eight minutes. The light from the stars of our Galaxy travels for tens to hundreds of years, and light takes millions and hundreds of millions of years to travel to us from distant galaxies and clusters of galaxies. When an object is farther away, we can see it at an earlier epoch. The farthest objects are quasars, and we see them such as they were thousands of millions of years ago. While observing the distribution and motion of galaxies, their clusters, and superclusters, we learn their properties such as they were very long ago. However, a difference of hundreds of millions of years is not very large on the scale of the universe: its expansion occurs at a rate such that the density of matter at the present stage of the universe's expansion can only change considerably in thousands of millions of years. This is why the density of the nearby area of the universe, where galaxies can be observed, is uniform, the same everywhere.

However, if we could see at greater distances, i.e. into the distant past, we would evidently find that the density there (i.e. then) must be greater than that nearby (i.e. now). A picture taken in real rays would therefore show us the universe to be nonuniform in its density: the farther away from us, the denser it would be.

According to the general principle of Einstein's theory, space itself in such a picture should be nonuniform in its geometrical properties. Moreover, when taken in real rays, space is always finite in its volume, whatever the future of the cosmological expansion might be.

The point is that light reaching us from remote sources is shifted to the red: electromagnetic wavelengths increase and therefore their frequencies decrease. As we have mentioned above, this is due to the Doppler effect owing in this case to the cosmological expansion, or the relative motion of galaxies. According to the Hubble law, if a source is farther away from us, the velocity of its recession is greater, and therefore its spectrum lines shift more to the red. There exists a great but finite distance such that the wavelength of the arriving light proves to be infinite, and therefore the frequency of the light is reduced to zero, so the source becomes invisible to us. Consequently, the universe has a horizon, and observations are only possible within it. The volume of space accessible for observation proves to be finite for this reason, and the mass of matter within it is also finite. The distance to the horizon, which is at present about 15,000-18,000 million light years, is the path traversed by light during 15,000-18,000 million years from the beginning of the cosmological expansion to the present epoch.

The inference on the existence of the horizon does not depend on whether there are astronomical bodies bright enough to be able to send light to us from however great a distance. It is rather a fundamental phenomenon due to the fact that any waves (or rays) can only propagate for a finite distance in a finite lapse of time.

The horizon as the limit of vision is evidently also the limit for any exchange of signals, and therefore the limit for any causal relationships. Two events can be causally related, for instance, one is a cause and the other is its effect, only if both of them occur within the horizon.

The horizon expands along with the expansion of the universe; its radius increases as the path traversed by light during the time from the beginning of expansion, and each subsequent epoch includes more matter within the horizon. The parts of the universe within the limits of the present-day horizon, which is occupied by the visible galaxies, were earlier separated from each other by the horizon. The horizon separated them as an impenetrable wall, and they "knew nothing" about each other then. However, coming into the horizon now, they "discover" that their densities are identical. Why so? Why do the distribution and density of matter prove to be correlated in these different and remote parts of the universe? There has been no suitable answer to this question so far. One can only assume that the universe was uniform "from the very beginning", and the cosmological expansion in each of its areas always occurred at the same rate. But what does it mean "from the very beginning", and what was "before"? These are the questions which appear to be harder to answer than those concerning the causal relationships in the universe.

Relict Radiation

Proceeding from the general laws of physics and Friedmann's theory of the cosmological expansion, G. A. Gamow assumed in the 1940's that very long ago the universe was both very dense and very hot. This idea was strongly supported by observations in 1965 when A. Penzias and R. Wilson detected microwave background radiation. The evidence showed that the entire universe is filled with radio waves in the millimetre range, propagating uniformly in every direction.

The quantum theory has long since established that any electromagnetic radiation has both wave and corpuscular properties, thus reconciliating the opponents in the long-standing dispute. Cosmic radio waves can also be regarded as a population of particles, quanta of electromagnetic radiation called photons. This "photon gas" uniformly fills the entire universe. The temperature of the photon gas is close to absolute zero: it is about three kelvins. But the energy it contains is greater than the light energy radiated by all stars combined during their life spans. There are about 500 quanta of radiation per each cubic centimetre of the universe, while the total number of photons within the limits of the visible universe is several thous and million times greater than the overall number of matter particles, i.e. atoms, nuclei, and electrons of which planets, stars, and galaxies consist.

This general background radiation in the universe has been called by I. S. Shklovsky "relict radiation", and this is in fact a relic that survived from the hot and dense initial state of the universe. The possibility of discovering relict radiation against the background of the radiation from galaxies and stars in the centimetre radio wave range was substantiated in the calculations performed by A. G. Doroshkevich and I. D. Novikov in 1964, a year before the discovery made by A. Penzias and R. Wilson.

It is well known that heated matter always emits photons. According to the general laws of thermodynamics, this reveals the tendency to the equilibrium state at which saturation, as it were, is attained: the birth of new photons is compensated for by the reverse process, i.e. the absorption of photons by matter, so the overall number of photons in the medium does not change. When this thermodynamic equilibrium of matter and radiation is attained, the photon average energy is proportional to the temperature, while the concentration of photons, i.e. their number per unit volume, is proportional to the cube of the temperature.

Assuming that the matter in the early universe was hot, G. A. Gamow predicted that the photons which were then in the thermodynamic equilibrium with matter would have remained till the present time. These photons were discovered in 1965. Having been subjected to the general expansion and the related cooling, the photon gas became the background radiation of the universe, uniformly arriving from every direction.

Just as any quantum of electromagnetic radiation, a quantum of relict radiation does not possess any rest mass, but it has certain energy and therefore mass related to this energy in keeping with Einstein's famous formula $E=Mc^2$. This mass is very small for most of these relict quanta, much less than the mass of an atom of hydrogen, the most abundant element in stars and galaxies. Therefore, despite the considerable predominance in the number of quanta, the contribution of relict radiation to the total mass of the universe is less than that of stars and galaxies. At present, the density of radiation amounts to 3 × 10^{-34} g/cm^3, which is about a thousand times less than the average density of matter in galaxies.

But this has not always been so: there was a period in the history of the universe when photons made the major contribution to its density. The point is that during the cosmological expansion, the density of radiation declined faster than that of matter. The concentration of photons decreased in the process (at the same rate as the concentration of material particles), but the average energy of a photon declined as well because the temperature of the photon gas dropped during the expansion. The density of radiation energy is the product of the average energy of a quantum and the concentration of quanta; the mass density of radiation is derived by dividing the density of energy by the square of the velocity of light. The density of radiation during the first million years of the cosmological expansion exceeded the density of matter (Fig. 6). During this early epoch, which can naturally be called the epoch of radiation prevalence, the gravitation of the cosmic medium and therefore its entire dynamics were controlled by radiation. The density of the universe was about 3× 10^{-22} g/cm^3 at the midpoint between the epoch of radiation prevalence and the epoch of matter prevalence.

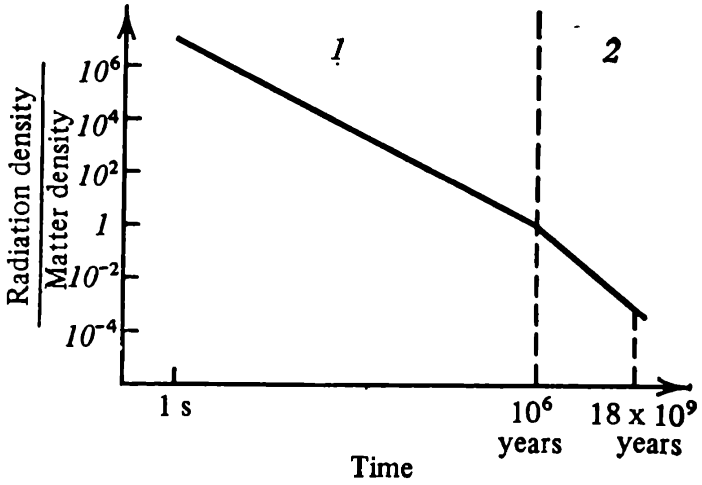

Fig. 6
The ratio of the radiation density to the matter density in the expanding universe. *1* — the epoch of radiation prevalence; *2*—the epoch of matter prevalence. Logarithmic scales along both axes.

Cosmological theory gives a law according to which every distance and length in the universe change. First, distances and lengths increase proportionally to the square root of time, but then, past the moment of equal density of matter and radiation, they increase faster, as time to the power 2/3. During both these epochs, the cosmological expansion occurs with a deceleration due to the gravitation of the medium itself. The recession velocities of cosmic bodies would evidently be constant without this deceleration; but then distances and lengths would change in proportion to time. However, the actual expansion is always slower than in direct proportion to time.

Imagine a sphere in the universe filled with some given particles. The sphere's radius will increase with time according to the law mentioned above. The total mass of the particles (more accurately, their rest mass) does not change during the cosmological expansion. But the total mass of photons, i.e. the total mass of radiation, decreases in inverse proportion to the radius of the sphere. The density of matter drops in inverse proportion to the volume of the sphere, i.e. the cube of its radius. The concentration of the photons within the sphere varies according to the same law. Therefore the ratio of the mass (density) of radiation to the mass (density) of matter decreases during the expansion in inverse proportion to the radius of the sphere and generally to every length and distance in the universe. This is shown in Fig. 6.

During the entire epoch of radiation prevalence, the matter in the universe was so hot that thermal motion prevented electrons from joining their nuclei. Neutral atoms could not exist, and the cosmic medium was in the state of complete ionization, i.e. the state of plasma. During the first instants of the cosmological expansion, the first fractions of a second after its beginning, the temperature of the medium exceeded 10^{11} kelvins, and the thermodynamic equilibrium was reached in this hot "cauldron", and the currently existing ratios between quanta and material particles, the components of the cosmic medium, were established. Besides photons, there also existed the neutrino gas in this equilibrium with matter, the number of neutrinos being almost the same as that of photons. Just as photons, neutrinos were to survive till the present epoch, and although they cannot be registered directly, there is no doubt that relict neutrinos really exist. According to a hypothesis, the "hidden mass" revealing itself in the dynamics of the motion of galaxies and their clusters consists of neutrinos.

Neutral atoms could not exist at the high temperatures mentioned above, and composite atomic nuclei could not exist either; the environment was a mixture of elementary particles such as neutrons, protons, electrons, neutrinos, photons, etc. Thermal motion did not allow neutrons and protons to join and become bonded by nuclear forces, and accidental nuclei were immediately split by colliding particles. The temperature was falling during the expansion, and the formation of nuclei became possible 3-5 minutes after time zero, when the temperature declined to 10^9 kelvins.

The theory of the formation of chemical elements in the early universe, also called the theory of initial nucleosynthesis, is a major achievement of cosmology. Most protons remained free, and the protons which did not form complex nuclei determined the overall spatial concentration of hydrogen, the lightest element, the nucleus of whose atom is a proton. Hydrogen amounts to 70-75 mass per cent of the universe. Helium, the element following hydrogen in Mendeleev's periodic table, took almost all protons left and the same number of neutrons (there are two protons and two neutrons in the nucleus of a helium-4 atom, the principal helium isotope). Helium amounts to about 25-30 mass per cent of the universe. Other nuclei, those of heavy isotopes of hydrogen (deuterium and tritium), the light helium isotope (He-3), and even nuclei of such light elements as lithium, beryllium, and boron, were produced in quantities of no more than several hundredths of one per cent. Heavier nuclei could not be formed in this process which lasted about a quarter of an hour. They appeared much later, in nuclear reactions within stars.

The ratio between hydrogen and helium indicated by the cosmological theory of element formation is in close agreement with astronomical data on the chemical composition of the atmospheres of the oldest stars where, as we may assume, the "primary" matter of which these stars were formed prevails.

The first calculations of cosmological nucleosynthesis were performed towards the late 1940's, long before the discovery of relict radiation. The agreement of the results on the cosmic abundance of hydrogen and helium made it possible to predict at that time the expected temperature of the present-day relict photons; Gamow's estimates showed that it should be between one and ten kelvins above absolute zero. The actual temperature, three kelvins, is exactly within this range.

More details on the diverse physical processes during the earliest stages of the cosmological expansion canbe found in the book by S. Weinberg listed in Recommended Literature.

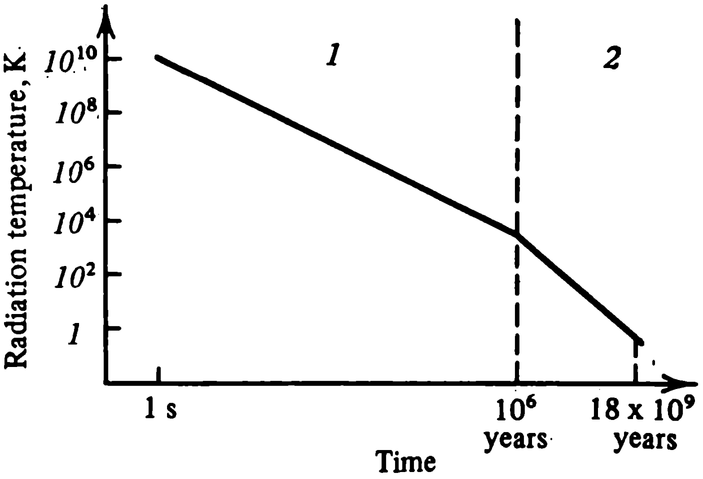

Fig. 7
The temperature of radiation in the expanding universe. During the epoch of radiation prevalence (*1*) it coincides with the temperature of matter. Logarithmic scales along both axes. This graph is very similar to Fig. 6 (see text).

After the formation of helium and a small admixture of other light nuclei, the evolution of the cosmic medium occurred without variation in its composition, and only the ratio between the densities of radiation and matter gradually changed in favour of the latter. The temperature also dropped, declining in inverse proportion to all lengths and distances, i.e. according to the law governing the ratio of the density of radiation to that of matter (Fig. 7). When the age of the universe was about one million years, the temperature decreased to 3000 kelvins, and then the recombination of cosmic plasma occurred: electrons and protons, previously separated by thermal motion, joined each other, forming neutral atoms of hydrogen. The recombination of helium had taken place somewhat earlier: each helium nucleus joined two electrons and formed a neutral atom of helium. The transformation of the ionized medium, the primordial hydrogen-helium plasma, into a gas of neutral atoms was coincident (it is yet unknown whether accidentally or not) with the transition from the epoch of radiation prevalence to the epoch of matter prevalence.

Before recombination, radiation and ionized matter interacted electromagnetically and were in a thermodynamic equilibrium. After recombination, the interaction between radiation and neutral matter was no longer possible. The state of equilibrium between them also came to an end, and matter, which was cooling faster than radiation during the cosmological expansion, later possessed a lower temperature than that of the photon gas. It was only after recombination that the remaining photons turned into relict photons.

Following the recombination epoch, the formation of the observed cosmic structures began in the universe. We are going to deal with these structures in the next chapters.

Chapter 2

The Origin of the Large-Scale Structure of the Universe

Neither stars nor galaxies existed at the high temperatures and densities of the early universe. Even atomic nuclei heavier than a proton could not exist at first, and atoms appeared only during the recombination epoch, one million years after the beginning of the cosmological expansion. It can be said that "from the very beginning" the universe had no "bonded" formations, of either astronomical scale or microscopic scale, i.e. that of atomic nuclei.

However, the complete absence of structure, ideal uniformity and isotropy on all scales were impossible. Some deviations from uniformity and isotropy always existed and involved considerable masses of matter corresponding to those of the observed cosmological systems, from stars to the largest clusters and superclusters of galaxies. This was the prestar, pregalactic structure of the universe. Its elements were not isolated condensations of matter, but rather some weak and shapeless irregularities of the medium, which physicists call small perturbations.

This chapter deals with the prestar, pregalactic structure of the universe, its evolution throughout the general cosmological systems, i.e. clusters and superclusters of galaxies.

Stars, Galaxies, and Cosmological Expansion

Current concepts of the appearance of astronomical structures proceed basically from the cosmological theory defining the general pattern of the development of the universe as a whole. Cosmology is abstracted from such "details" of the universe as stars, galaxies, and even clusters and superclusters of galaxies, regarding them as randomly moving "points" of which the metagalactic medium consists as gas consists of molecules. The average density of the metagalactic medium is uniform, i.e. it is the same at characteristic lengths exceeding 100-

300 Mpc; this follows from counting stars and galaxies in sufficient volumes of space: the number of stars or galaxies in any volume of large enough size is the same wherever the volume is selected.

Cosmology points out that all matter in the universe expands. This is only relevant to large volumes, 100-300 Mpc (and larger) in size. But within these volumes, stars and galaxies "do not notice" the cosmological expansion. No expansion related to the general expansion of the world occurs in stars and star systems; the same applies to our solar system.

Astronomical objects, from planets to clusters of galaxies, "do not remember" that long ago their matter was much denser and hotter, that it was then uniformly mixed in space, and that it expanded as everything else. The formation of astronomical systems required the suppression of this general expansion in certain volumes of the medium. The only force capable of performing this was the gravitation produced by the very matter of the universe.

The force of gravitation tends to bring bodies or material particles closer—always and everywhere. It operates throughout the universe, and this is why it has been called universal since Newton's time. We do not know why the general recession of cosmological systems occurs, but there is no doubt that since the first instants of the cosmological expansion, universal gravitation interfered with this expansion and strove to suppress it. Gravitation failed to stop the expansion of the universe: the initial push was too strong; but gravitation did succeed in limited regions of the universe, although they are very great in size and mass.

The age of our Galaxy as a star system is close to the age of its oldest stars estimated at 10,000-12,000 million years. Evidently, the process of the formation of the observed astronomical bodies started 12,000-15,000 million years ago, about 1000-3000 million years after the beginning of the expansion.

But what did it begin with: the smallest or the largest bodies? There are many indications that huge masses of matter commensurate with those of clusters and superclusters of galaxies were the first to drop out from the cosmological expansion. Then the process of the fragmentation of these masses began and the entire hierarchy of astronomical systems was gradually built up within them. This is Ya. B. Zeldovich's viewpoint. It gives a formation pattern of the large-scale structure of the universe, which appears to be the most detailed, elaborate, and convincing. We shall deal with it below.

Another viewpoint assumes the initial appearance of smaller bodies which later joined each other and formed larger bodies. This idea is upheld by J. Peebles and R. Dicke. They suggested that the first objects in the universe could have been bodies of about a million Sun masses. Step by step, joining each other, they formed galaxies, and the union of galaxies produced their clusters. Some of the original bodies are preserved and appear as the globular star clusters known both in our Galaxy and in other galaxies.

Gravitational Instability

That the gravitation of the medium itself played the principal role in the formation of astronomical bodies was felt long before modern cosmogonical research. Isaac Newton was the first scientist to indicate this.

On 10 December 1692, Isaac Newton wrote to Richard Bentley, the rector of the Trinity College at Cambridge,

"It seems to me, that if the matter of our sun and planets, and all the matter of the universe, were evenly scattered throughout all the heavens, and every particle had an innate gravity towards all the rest, and the whole space throughout which this matter was scattered, was finite, the matter on the outside of this space would by its gravity tend towards all the matter on the inside, and by consequence fall down into the middle of the whole space, and there compose one great spherical mass. But if the matter were evenly disposed throughout an infinite space, it could never convene into one mass; but some of it would convene into one mass and some into another, so as to make an infinite number of great masses, scattered great distances from one to another throughout all that infinite space. And thus might the sun and the fixed stars be formed, supposing the matter were of lucid nature."

In this famous cosmogonical outline, Newton divided the problem into what is subject and what is not subject to scientific research and explanation. The problem why stars (the luminous matter) are luminous and planets (the dark matter) are not was considered by him to be not subject to research. This was really a difficult problem, and it was only two and a half centuries after Newton that the source of luminosity of stars and the Sun was finally revealed and explained. It was necessary to develop nuclear physics along with quantum mechanics and the special theory of relativity to understand this in full.

However, the first aspect of the many-sided problem, the formation of celestial bodies from uniform matter, was revealed by Newton in principle. The very idea of the universe evolving owing to natural physical laws is remarkable. Today we can directly observe the evolutionary processes in astronomy on the surface of the Sun, in comets, stars, pulsars, and the nuclei of galaxies. However, in Newton's time there were no real indications of anything of the sort. The idea and the mechanism of cosmological evolution were born out of Newton's law of universal gravitation, and they proved to be valid for the entire development of cosmogony. Gravitation shaped the cosmological structure, and its further complication and differentiation gave rise to an evolution of another kind, the result of which was that matter existing 12,000-15,000 million years ago in a state of uniformly distributed plasma reached higher level of organization, and life appeared in the universe, and then the human intellect evolved, which was capable of studying its cosmological prehistory, among other things.

The mechanism of the formation of celestial bodies from uniformly distributed matter operates, according to Newton, owing only to gravitation; no other forces, if any, can prevent gravitation from exerting its influence. Considering uniform matter, there should be no forces capable of counteracting gravitation. The force of pressure, for instance, appears only where there is a nonuniformity, a drop in pressure between two sites. Gravitation in a star is actually balanced by pressure because a star is nonuniform: the pressure in its centre is greater than that at the surface. However, any uniform gravitating medium cannot be at rest: this is Newton's original idea. A medium cannot be at rest when there is an unbalanced force of gravitation. Such a state is unstable.

The first attempt to develop Newton's idea was made by J. Jeans towards the early 20th century. In his book *Astronomy and Cosmogony*, Jeans published Newton's famous cosmogonical letter (the above quote is from this book) and explicitly indicated that the key mechanism of the formation of celestial bodies is the gravitational instability of a uniform medium.

This process is easy to imagine. Suppose the density at a certain site of a uniform medium becomes somewhat greater. Then the gravitation at this site increases as well. The gravitation tends to draw the particles of the medium closer together, and therefore to increase the density again. Because of this new increase in the density, the gravitation becomes stronger, etc. Once started, such a process continues and develops by itself, causing ever increasing deviations from the initial state. This phenomenon is called instability.

J. Jeans noticed that a factor impeding the development of gravitational instability is the elasticity of a medium: a contraction at a given site of the medium brings about greater pressure, and the force of pressure, appearing because of the difference between the pressure in the condensation and its vicinity, strives to expand the condensation. But the force of pressure is less than the force of gravitation if the area of contraction is large enough. This is natural because the force of gravitation is greater, the greater the mass (and therefore the size) of the condensation. The force of pressure is less, the greater the length at which a given drop of pressure is felt.

The critical size at which the forces of gravitation and pressure are comparable is called the Jeans length. This length reduces as the density of the medium increases, while it increases as the pressure of the medium grows. This kind of dependence on pressure and density is natural: pressure reflects the elasticity of the medium, its ability to withstand contraction, while the force of gravitation depends on the density of the medium. The elasticity of a medium can also be characterized by the velocity of sound in it; the Jeans length is directly proportional to the velocity of sound. And most commonly the critical length is expressed through the velocity of sound: $u/\sqrt{G\rho}$. Here u is the velocity of sound, ρ is the density of the medium, and G is the gravitational constant (recall its value: $G=6.7\times10^{-8}$ cm^3/(g·s^2)).

For instance, if the density of a medium is 3×10^{-22} g/cm^3 and the velocity of sound in it amounts to 6 km/s (these were the conditions in the universe during the recombination epoch and we shall come back to them below), then the Jeans length equals 2×10^{20} cm. An area of such dimensions at the indicated density contains a mass approximately equal to a million Sun masses.

Along with the Jeans length, we shall often resort in our considerations to the Jeans mass, i.e. the mass of matter in a sphere of radius equal to the Jeans length.

The idea of gravitational instability was reestablished in Jeans' theory both too late, two centuries after Newton, and too early, before Friedmann's cosmology. In Jeans' theory, limited volumes of a uniform medium contract; there may be a however great number of such volumes, and the behaviour of each of them does not depend on the others. But the entire infinite uniform medium is at rest. It is quite obvious that such a rest is infeasible: the force of gravitation acts upon the entire medium, and this force is not balanced by anything.

A consistent theory of gravitational instability was developed in 1946 by E. M. Lifshits proceeding from Friedmann's cosmology. This theory treats thoroughly the behaviour of weak contractions and generally any small perturbations in a uniform expanding medium. The discovery of relict radiation introduced new essential components in Lifshits' theory, and currently this theory is a reference point for all researches, hypotheses, and cosmogonical schemes worked out to explain the nature of the large-scale structure of the universe.

Gravitational instability operating in a uniform expanding medium reveals itself in that separate volumes whose densities prove for some reason a little greater than the general density of the medium expand more slowly than the medium as a whole. This is evidently caused by gravitation, which is stronger in such volumes and therefore more effectively impedes the expansion. Although the densities of both perturbed volumes and the entire medium decline with time, the perturbed volumes lag ever more behind the general expansion, and therefore the difference in the densities increases (Fig. 8).

According to Newton and Jeans, the process developed from the state of rest, and the denser volumes contracted; however, the denser volumes in the expanding medium continued to expand. In both cases, the original weak deviation from the general density of the medium increased spontaneously. It can be said that in both cases the matter in denser volumes acquired proper motions and proper velocities; the denser volumes in Newton's picture contracted, and the proper velocity is the velocity of contraction. Proper motions in the expanding medium are the differences between the true motions in the denser volumes and the regular expansion of the entire medium, and any proper velocity is the difference between the velocity of the general expansion and the true velocity in the given volume. Proper motions are opposite to the general expansion. The gravitational instability in the expanding universe intensifies weak perturbations of density and the respective proper motions.

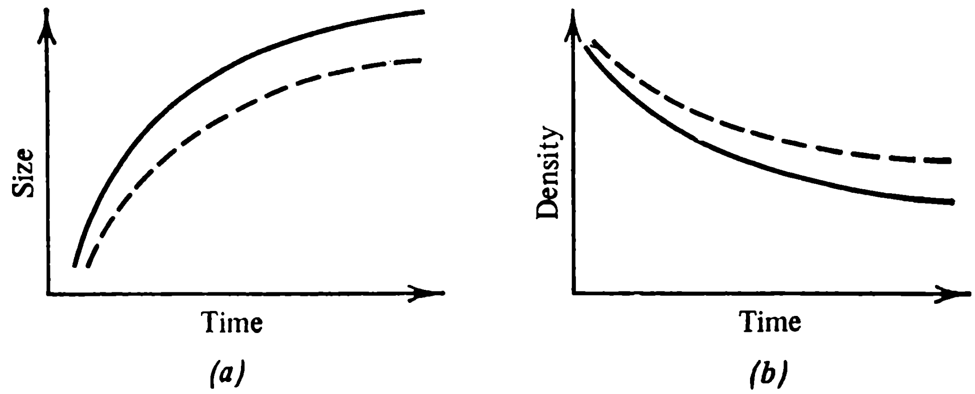

Fig. 8
Gravitational instability in the expanding universe. (*a*) The size of a perturbed volume (dotted line) increases more slowly than the size of the identical volume in an "unperturbed" medium (solid line). (*b*) The density of a perturbed volume (dotted line) declines more slowly than the density of the medium as a whole (solid line).

The denser volumes gradually lag behind in their expansion, it is impeded, and sooner or later there is an instant when the expansion in these volumes halts and therefore they "drop out" from the general cosmological expansion. As to proper motions, the halt means that the proper velocities of the denser volumes become identical to the velocity of the regular expansion. This implies that weak perturbations transform into strong perturbations. New interesting processes come into play, and we shall deal with them below.

Lifshits' theory makes it possible to determine how much time it takes for a given perturbation to transform from a weak into a strong one. Density perturbations are usually characterized by a relative value, i.e. the difference of densities in the perturbed volume and the entire medium divided by the density of the medium. This relative value of density perturbations increases in proportion to time during the first million years after the beginning of the expansion, and later (after the recombination epoch) in proportion to time to the power 2/3. While the age of the universe grew within the first million years, for instance, a thousand times, the relative value of density perturbations increased a thousand times as well, but later the same duration of time only resulted in a hundred times gain in density perturbation.

A density perturbation becomes strong when its relative value approaches unity, i.e. when the density in a condensation is about twice that of the medium. According to what has been presented at the beginning of this chapter, the epoch of strong perturbations is the time of the separation of the first condensations corresponding to the world age of 1000-3000 million years. It is easy to calculate that during the time from the age of one second (i.e. from the epoch starting from which we can reliably use ordinary physics to describe the universe) to 3000 million years, the relative value of perturbations increased no more than 10^{17} times.

This increase implies that when the world was one second old, the metagalactic medium should have possessed perturbations whose relative value was no less than 10^{-17}. Therefore the theory indicates the condition of weak "triggering" perturbations existing in the universe long ago; it suggests the degree to which the cosmic medium was perturbed during the epoch. And we see that density perturbations were very weak indeed when the universe was one second old.

However, they were not too small: compare these perturbations with thermal fluctuations in a medium. Thermal fluctuations always occur when the temperature of a medium is not absolute zero. They appear because random thermal motion of particles (recall Brownian motion) can accidentally lead to an increase in the density of particles in some volumes of the medium and a decrease in others.

Note that in Jeans' theory it was precisely thermal fluctuations of the medium that were given the role of the initial triggering of gravitational instability.

Proceeding from the general relationships of thermodynamics, we can find the characteristic values of density fluctuations encompassing a given number of particles. These characteristic values do not depend on temperature (it is only required that it is above absolute zero) and are only expressed through the number of particles involved in the fluctuation. The relative value of such fluctuation density perturbation equals the reciprocal of the square root of the number of particles. Suppose we are interested in a fluctuation involving a number of particles such that is contained in a galaxy. There are about 10^{68} particles in our Galaxy. The relative value of fluctuation density perturbation for this number of particles amounts to 10^{-34}. As we see, this is many times less (refined calculations only increase the difference) than the perturbations which should have existed in the universe at the age of one second.

Hence a very essential conclusion follows: the initial perturbations intensified by gravitational instability should be much greater than the level of thermal fluctuations during the entire period of the universe's evolution, which can be studied on the basis of the established physical laws. This implies that we can deal with these perturbations as a prestar, pregalactic structure existing in the universe long before stars and galaxies appeared.

The problem of this original structure is one of the most difficult in cosmology. It is currently treated on the basis of the most advanced ideas unifying the concepts of both the general theory of relativity and the quantum theory.

The origin of initial perturbations is most likely due to the same processes which brought about the cosmological expansion itself; however, we have learned too little about it so far.

It is noteworthy that we can make definite conclusions on the pregalactic structure which existed as early as the first seconds after the beginning of the cosmological expansion. The further evolution of the structure, up to the stage of strong perturbations, can be studied in appropriate detail.

Pregalactie Structure

The modern theory of gravitational instability operates with the critical Jeans length, just as the initial Jeans' theory did. The Jeans length even retained its expression through the velocity of sound and the density of the medium (see above). This is understandable: the Jeans length appears in the considerations of the role of pressure and gravitation in the medium, and these considerations are so simple and natural that they do not require a knowledge of the general picture of gravitational instability.

Very essential is that the relationship between the size of perturbed volume and the Jeans length in an expanding medium changes with time. We shall see below that the Jeans length increased in proportion to time from the beginning of the expansion during the epoch prior to the recombination of cosmic plasma. However, the size of perturbations only increased in proportion to the square root of time, as follows from the general law governing distance changes in an expanding universe. (The size of the condensations intensified by gravitational instability increased even more slowly because the expansion of the condensations lagged behind the general cosmological expansion.) While the size of perturbations during some earlier period exceeded the Jeans length, it could have become less than the Jeans length later. Then gravitational instability in this volume would have ceased, and theory shows that the perturbations with characteristic lengths less than the Jeans length transformed into pulsating condensations and rarefactions similar to sound waves in a medium. The amplitude of such waves did not vary during the course of the general expansion (prior to the recombination epoch), whereas the wavelengths increased.

The cosmic medium during the first million years of the expansion was a mixture of electromagnetically interacting plasma and radiation. Photons prevailed in both number and density. The number of photons in the universe is approximately a thousand million times greater than that of electrons and protons. Photons do not have any rest mass, and therefore their "rest density" equals zero. However, they possess energy, and energy is always related to a certain mass. This is because Einstein's formula $E=Mc^2$ associates any mass M with an energy E, and vice versa, any energy E is associated with a mass M. Mass and energy are directly proportional, the proportionality factor being the square of the velocity of light.

Radiation prevailed in this mixture of cosmic plasma and radiation, and the elasticity of the medium and the pressure in it depended on radiation. In agreement with this, the velocity of sound in the medium was controlled by radiation, i.e. photons. The velocity of sound in a common gas is almost identical to the average velocity of molecular thermal motions. The thermal velocity in the photon gas is therefore the velocity of light with which every photon moves. This is why the velocity of sound in the photon gas (with a small admixture of plasma) is close to the velocity of light: $u=c/\sqrt{3}$.

Knowing the velocity of sound, we can estimate the Jeans length according to the general relationship mentioned above: the Jeans length is $u/\sqrt{G\rho}$. Here ρ implies the total density of the cosmic medium, i.e. the total of the densities of matter and radiation. We also know that there is a simple relationship between the density ρ of the universe at a given moment and the time t from the beginning of the cosmological expansion to this moment: $t \approx 1/\sqrt{G\rho}$. Comparing the two relationships, we can see that the Jeans length in the early universe increased in proportion to the age of the universe. Within the accuracy of an inessential numerical factor, this length is the product of the velocity of light and the time from the beginning of the cosmological expansion to the given moment: *ct*.

We can see that the Jeans length in the early universe is identical (within the same accuracy) to the distance to the horizon of the universe: the distance to the horizon is the path traversed by light during the time from the beginning of the cosmological expansion to the present. Growing perturbations in the hot universe of the first million years reach beyond the horizon. There can be no causal relationship between the boundaries in a condensation of such a size: a boundary "does not know" what occurs at the opposite end. But gravitational instability acts at the same rate within the entire volume of a condensation, and therefore enhances each perturbation as a whole. In fact, this situation is the same as in the entire universe: the size of the universe is greater than the distance to the horizon, there is no causal relationship between the universe and the areas beyond the horizon, and nevertheless the entire universe expands at a regular rate identical in all its areas.

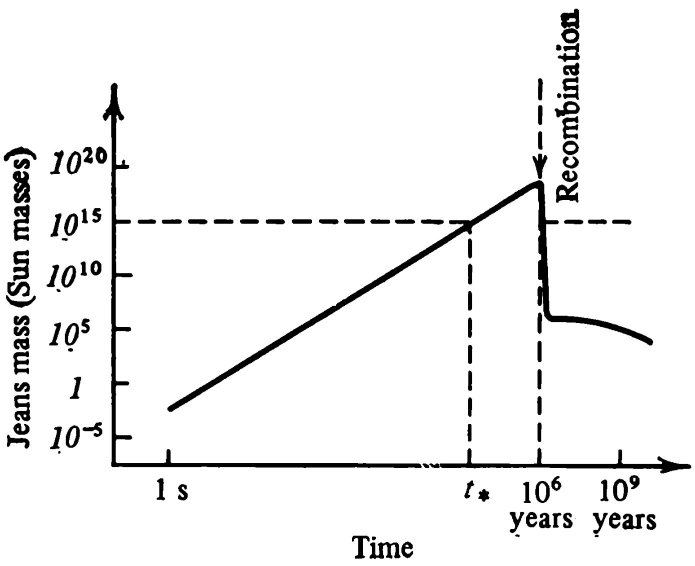

Fig. 9
The Jeans mass in the hot expanding universe. Logarithmic scales along both axes.

To investigate gravitational instability in the early universe, it is essential to know how the Jeans mass, i.e. the mass of matter within a volume of the Jeans length, varies from one instant to another. The temporal variation of the Jeans mass is presented in Fig. 9. This mass was relatively small, less than the Sun mass, when the universe was one second old. However, it increased rapidly during the expansion, and when the universe was one million years old, it amounted to an immense value of 10^{18} Sun masses, which is a thousand times greater than the mass of the greatest cluster of galaxies. The Jeans mass implies here the mass of matter or the mass of plasma particles within a volume of the Jeans length radius; it does not include radiation: radiation would be gone, and a galaxy or a cluster of galaxies would only be formed out of plasma particles.

Any perturbation encompassing as many plasma particles as there are in a major cluster of galaxies with a mass of 10^{15} Sun masses exceeded the Jeans length at almost any time (see Fig. 9).

The recombination of plasma occurred when the universe was one million years old, and it resulted in an abrupt change of physical conditions in the metagalactic medium. Prior to this, matter had been in the state of plasma, and electrons had been separated from ions, mainly protons, by thermal motions; however, the general cooling of the medium during the cosmological expansion gradually weakened the thermal motions. Finally, the temperature dropped to 3000 kelvins, and electrons could join ions and form neutral atoms. It is very essential that recombination stopped the interaction between radiation and matter, and that matter became neutral.

After recombination, radiation and matter behaved independently. The condensations of matter continued to become denser because of gravitational instability, but the Jeans length corresponding to the new conditions and the respective Jeans mass dropped abruptly (see Fig. 9). This happened because the proper elasticity of the gas rather than photons controls the velocity of sound in the gas. At a temperature of 3000 kelvins, the velocity of sound is about 6 km/s. The density of the universe during the recombination epoch is estimated at 3×10^{-22} g/cm^3. We have already given an estimate of the Jeans mass for such values of the velocity of sound and the density of the medium and found it equal to a million Sun masses. The drop in the Jeans mass after recombination from 10^{18} to 10^6 Sun masses is shown in Fig. 9.

During the post-recombination epoch, perturbations in masses exceeding a million Sun masses were caused by gravitational instability at the same rate. Figure 10 shows the temporal variation of the relative value of density perturbations on the scale of 10^{15} Sun masses, corresponding to a cluster of galaxies.

During the period when gravitational instability was inoperative on the given scale, condensations exhibited, as it has been mentioned, oscillations of density, whose amplitude was constant. This is shown in Fig. 10.

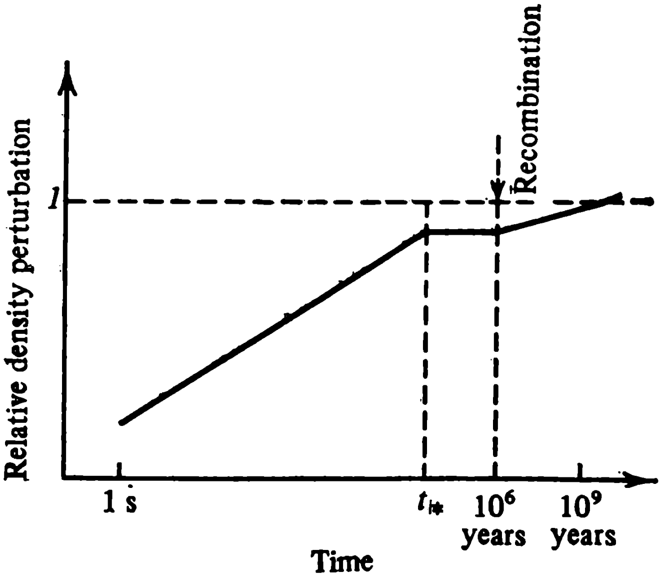

Fig. 10
The growth of perturbations in the hot universe. The relative density perturbation grows with time if the mass of a perturbation exceeds the Jeans mass. If the mass is less than the Jeans mass, the density perturbation does not increase; it oscillates with a constant amplitude. The damping of perturbations on the scales less than the scale of the largest clusters of galaxies is disregarded (see text). Logarithmic scales along both axes.

Large-scale perturbations have the advantage of virtually not being subject to dissipative processes. Dissipative processes are related to the viscosity and thermal conduction of the medium, tending to smooth out any perturbations in the medium and convert the related energy of proper motions into heat. These processes do not affect the general cosmological expansion because the medium remains uniform, and there is no sliding of layers onto other layers. However, density perturbations can be subject to their effect. Heat conduction is very essential during the early epochs of the cosmological expansion.

During the prerecombination epoch when matter and radiation were closely interrelated, they behaved as a single medium. Both plasma particles and radiation photons participated in density perturbations intensified by gravitational instability. These joint perturbations of plasma and radiation are called adiabatic perturbations. The density and temperature of adiabatic condensations were greater than those of their environment. But a drop in temperature always induces a flow of heat from the hotter area into the cooler area. Heat transfer in a medium which was a mixture of plasma and radiation was carried out best by photons: they "leaked" from the areas of condensation, carrying the excess heat with them. The same occurred with plasma particles, but photon, or radiative, heat transfer was more effective: first, photons prevailed over plasma particles in both number and density; second, photons escaped from the areas of condensation more easily than electrons or ions. The escaping photons dragged electrons and ions along, and therefore the condensations dissipated.

J. Silk and G. V. Chibisov showed that this radiative heat transfer smoothed out the drops in temperature and also damped the adiabatic density perturbations in the cosmic mixture of matter and radiation. The less the characteristic length of a perturbation, the greater is this effect. This is related simply to the fact that the excess, extra photons left small-size condensations sooner than large-size condensations. The result was that, by the recombination epoch, when the action of the radiative heat transfer ceased, every adiabatic perturbation containing a mass of less than 10^{15} Sun masses had been eliminated.

This is a very significant result. It means that the adiabatic perturbations that survived after the first million years of the cosmological expansion of the universe were such whose mass corresponded to the major formations, i.e. clusters and superclusters of galaxies. The cosmogonical theory currently developed by Ya. B. Zeldovich and his colleagues is underlain by the assumption of the existence of primary adiabatic perturbations, and this fact is of key importance: the development of gravitational instability brought about the separation of condensations of 10^{15} Sun masses (and greater) during the post-recombination epoch.

Entropy Perturbations

And still, the "survival conditions" for a perturbation in the universe were not so severe as they might seem. The point is that joint perturbations of plasma and radiation (the adiabatic perturbations discussed above) belonged to one of the two types of weak density perturbations in the early universe. Perturbations of the other type involved only material particles: these are condensations and rarefactions of plasma against the background of completely uniform radiation. These are called entropy perturbations, and they were investigated during the 1950's by L. E. Gurevich and A. I. Lebedinsky in massive stars with great radiative pressure. In 1966, Ya. B. Zeldovich drew attention to the essential role of entropy perturbations in the cosmology of the hot universe.

Plasma nonuniformities were, as it were, "frozen" in the "unperturbed" background of radiation, and they neither dissipated nor were intensified during the entire first million years when matter remained ionized and closely interrelated with radiation electromagnetically.

Since the concentration of photons did not exhibit any perturbations and was the same both within and without condensations, there was naturally no flow of photons from condensations outwards, and therefore the radiative heat conduction did not take place. Gradual decay of plasma condensations only occurred in the smallest perturbations; but this is how condensations of about one Sun mass could disappear, while condensations of larger sizes remained perfectly unchanged by the recombination epoch. To be more accurate, it has to be said that the relative density perturbation, i.e. the ratio of the density excess in a condensation to the "unperturbed" density of plasma, remained unchanged, while each condensation as a whole participated in the general cosmological expansion and its dimensions increased just as every other dimension in the expanding universe did.

The recombination put a stop to the interaction between matter and radiation. The earlier entropy perturbations yielded condensations of neutral atoms, and these condensations could then be intensified by gravitational instability without hindrance. It was only necessary that the size of a perturbation exceed the Jeans length. As has been mentioned above, the Jeans length reduced rapidly during the recombination epoch, and a volume of comparable size contained about a million Sun masses.

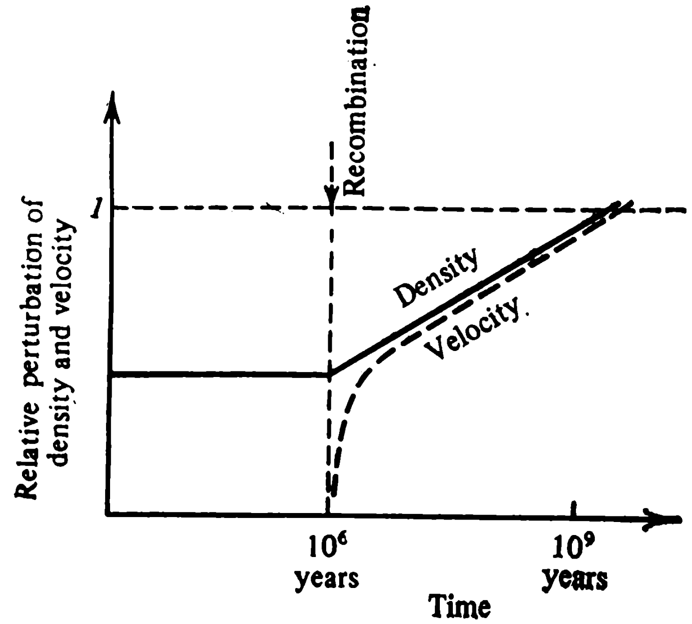

Fig. 11
The relative density perturbation and the relative proper velocity in entropy perturbations. Logarithmic scales along both axes.

Gravitational instability controlled these perturbations as it always had: it tended to decelerate the expansion. It communicated proper velocities to perturbations, and they were directed against the velocity of the cosmological expansion. These proper velocities increased with time, and therefore the condensations lagged behind the general expansion of the medium at an ever grêater pace.

Proper motions were superimposed, as it were, on the regular general expansion; they could be characterized by a relative value, the ratio of the proper velocity of a given perturbation to the regular velocity of the expansion within the same volume. The behaviour of the relative density perturbation and the relative velocity of the proper motion of a condensation brought forth by the initial entropy perturbations is shown in Fig. 11. Prior to the recombination, the relative density perturbation was unchanged, while condensations did not possess any proper velocity. After the recombination, density perturbations grew and proper velocities appeared. Both relative values varied according to the same regularity (in proportion to time to the power 2/3) after the initial rapid increase in the velocity to comply with density perturbations.

In 1968, J. Peebles and R. Dicke assumed that the entire observed structure of the universe appeared from original entropy perturbations. Proceeding from this, they developed a rather detailed hypothesis presenting the sequence of events from the recombination epoch to the recent epoch. They assumed that the relative density perturbation on the average increased as the mass of a primary condensation of plasma "frozen" into background radiation decreased. The dependence of the size of perturbations on their masses or characteristic length is called the perturbation spectrum; if perturbations on larger scales are weaker than on smaller scales, they are said to have a falling spectrum.

The falling spectrum of perturbations in the hypothesis of J. Peebles and R. Dicke implies that the first condensations capable of separating and withdrawing from the general cosmological expansion could have had masses comparable to a million Sun masses. Taking into account their mass, these were the condensations least intensified by gravitational instability after the recombination; the initial value of density perturbation in them was greater than in any other large condensation. Gravitational instability turned these condensations into rather dense clouds of more or less regular spherical shape.

The further evolution of each cloud was accompanied by gradual cooling (predominantly owing to the excitation and radiation of hydrogen molecules, which appeared, although in small numbers, in neutral atomic hydrogen after the recombination). The cooling made possible the fragmentation of gas clouds into smaller condensations, or protostars. In the end, the clouds turned into a cluster of stars whose mass, shape, and size were similar to the present-day globular star clusters. According to J. Peebles and R. Dicke, this was the pattern of origination of the globular clusters in our Galaxy and in other galaxies.

(The processes of fragmentation and star formation are treated in detail in Chapter 4.)

Gravitational instability, owing to which globular clusters are formed, also acts on larger scales. It develops in the "gas" whose particles make up these star clusters. This is why globular clusters collect into galaxies, and galaxies are then accumulated into clusters of galaxies.

Observational data on the star composition of globular clusters indicate that they are the oldest formations in the universe. This was the starting point for the hypothesis that J. Peebles and R. Dicke offered. Although galaxies are different in their mass, size, and shape, the difference between the stars in them is much less. The star population of the spherical component of our Galaxy is similar to the stars of elliptical galaxies, while the population of the disk of the Galaxy is like the stars of irregular galaxies. The similarities between galaxies are also exhibited in that many of them contain globular star clusters. The globular star clusters in our Galaxy populate the spherical subsystem; there are about a hundred and fifty of them, and they are all similar. The globular star clusters in some nearby galaxies are very much like those in our Galaxy.

The universal distribution of globular star clusters and their "old age" were directly explained in the hypothesis offered by J. Peebles and R. Dicke. However, a number of other essential properties of stellar systems could not be explained within the same hypothesis. Primarily, it concerns such properties of galaxies as their typical mass and density. Furthermore, recently there have also appeared objections to the starting point of the whole hypothesis: in the opinion of astronomers investigating globular star clusters, these clusters vary in their properties. Thus, the abundance of heavy elements in their stars depends on the position of a cluster in the Galaxy, while the chemical composition of stars in globular clusters in other galaxies correlates with the luminosity and mass of these galaxies. Finally, although it is hard to expect every globular star cluster to be within galaxies, for many of them should have remained in the intergalactic space, this is not in fact the case.

Peebles' and Dicke's hypothesis is one of the modern cosmogonical hypotheses. Produced on the basis of astronomical data and physical theory, such an evolutionary scheme is capable of revealing the possible succession of events resulting in the observed cosmological structure. This is not the only hypothesis, but rather one of a number of them; each of these hypotheses takes an astronomical fact or a physical process as a key point, but the choice remains ambiguous. This also refers to the hypothesis of strong hydrodynamic motions we are going to treat in detail below (and which we support).

In 1919, J. Jeans wrote, "In the present state of our knowledge any attempt to dictate final conclusions on the main problems of cosmogony could be nothing but pure dogmatism." Astronomical knowledge has increased immensely since that time, and theoretical physics went through enormous upheavals, but nothing can cancel what J. Jeans said.

Returning to entropy perturbations, let us note that it is not at all necessary that they are associated with the falling spectrum as asserted by J. Peebles and R. Dicke. It is also quite possible to assume that the greatest values of density perturbations belonged to condensations corresponding in their mass to the largest astronomical systems, viz. clusters and superclusters of galaxies, rather than to the smallest-scale condensations. Then these major condensations should have been the first to separate in the form of the gigantic clouds later disintegrating into smaller fragments.

Initial Perturbations and the Relict Background

Having studied the behaviour of both adiabatic and entropy weak perturbations in the early universe, we should ask the question: What was the actual prestar, pregalactic structure? Were its elements adiabatic or entropy perturbations or maybe a certain combination of them?

It is impossible to answer these questions proceeding only from theoretical deductions. In principle, theory allows for very different versions of cosmogonical schemes, and the choice between them can only be made on the basis of observational astronomical criteria. This is why observational data on relict radiation are very essential.

Relict radiation "tore away" from matter during the recombination epoch, and therefore it carries information about the universe as it was at the end of the first million years after the beginning of the expansion. Registering relict photons, we "photograph" the early universe. We could hope to see on such a "photograph" the condensations and rarefactions and proper motions of matter which occurred during that epoch and later gave rise to the cosmological structures observed. In order to see all that, it is necessary to investigate the distribution of the intensity of relict radiation over the sky. J. Silk drew attention to this in 1968.

Relict radiation is amazingly uniform and isotropic: it is received uniformly from every direction. And still, some irregularities or variations are bound to occur; there was a time when radiation was strongly associated with matter, and irregularities in the distribution and motion of matter left their traces in radiation.

Both matter and radiation participated jointly in adiabatic perturbations; therefore it is natural to expect that adiabatic perturbations left their traces in radiation at least in the form of weak variations of its luminosity from one site to another, or, which is the same, from one direction in the sky to another. It can be calculated that perturbations on the scale of a galaxy cluster could create variations in the intensity of background radiation in directions differing by several minutes of arc.

The search for "traces" of pregalactic perturbations in relict radiation have so far yielded no results: the relict background appears completely structureless. The greatest sensitivity in the observations of relict radiation was reached during the period from 1977 to 1980, using the newly built Soviet RATAN-600 radio telescope (600 m in diameter). Yu. N. Pariisky and his colleagues reported that if there were some angular variations in the intensity of the relict background, they did not exceed several hundredths or even thousandths of one per cent.

It should be said that proceeding from the concept of adiabatic perturbations, theoreticians had predicted earlier that for angles of several minutes the variations should be tenth fractions of one per cent, i.e. such that would have been immediately noticed with the RATAN-600.

The null result of these observations does not mean that there are no angular variations in the relict background at all; the sensitivity of the instrument only sets the upper bound to their value. If observations become more refined, this bound is extended and becomes more informative, and thus this theory gains a new impetus for further development rather than becoming invalid.

It turned out that the strict observational bound could be satisfied by a theoretical scheme where the pregalactic structure had been produced by entropy perturbations rather than adiabatic perturbations. In contrast to adiabatic perturbations, entropy perturbations only concern plasma in the early universe where the radiation remained uniform. This could result in the relative weakness of the traces it was capable of leaving in the relict background.

It can be said that the observations of the relict background provide evidence that entropy perturbations were the "building blocks" of the prestar, pregalactic structure of the universe.

And still, the problem of the type of initial perturbations cannot be regarded as finally solved. The point is that the theoretical estimate of the expected level of variations in the relict background depends essentially on whether neutrinos possess any rest mass. This connection between cosmology and the physics of elementary particles should not seem strange. Neutrinos exist in the universe in almost the same quantities as photons, and the origin of the prevailing number of both neutrinos and photons is relict, i.e. cosmological. If the neutrino rest mass is not zero, be it thousands of times less than the electron mass, cosmic neutrinos make the greatest contribution to the universe's density and therefore control gravitational fields, on the scale of both clusters of galaxies and the universe as a whole. Then the evolution of adiabatic perturbations on the scale of clusters of galaxies can only be accompanied by very weak variations in the relict background that can be detected within the observational bounds of the RATAN-600.

The Formation of Clusters of Galaxies

Whatever the nature of weak initial perturbations in the early universe, sooner or later they transform into strong perturbations under the effect of gravitational instability. Strong perturbations, irrespective of their prehistory, are condensations of matter where the cosmological expansion has been overpowered by their own gravitation—completely or almost completely.

Presenting above Peebles' and Dicke's cosmogonical hypothesis, we described the possible evolution of strong perturbations, close to the Jeans length in size, during the post-recombination epoch. The forces of pressure and gravitation were, as it were, balanced in the condensations: when the size of a condensation was equal to the Jeans length, both these forces were balanced exactly, while when the size was slightly more than the Jeans length, gravitation prevailed but pressure was nonetheless essential. This is why any condensation acquires a more or less spherical shape to which a gravitating mass of matter always tends if the forces of gravitation and pressure balance each other as, for instance, in a star.

The evolution of an intense condensation, whose size was considerably greater than the Jeans length, was quite different. Let us consider, according to Ya. B. Zeldovich, the physical processes in a condensation with a mass comparable to the masses of the major clusters of galaxies. The size of such a condensation in the post-recombination epoch was, in fact, much greater than the Jeans length (see Fig. 9).

It is striking that the shape of a large condensation does not tend to become spherical at all: the force of pressure in it is too weak compared with the force of gravitation, and this hinders any tendency to a balance and an equilibrium spherical shape. Conversely, if a cloud separated from the general cosmological expansion was initially a sphere, it dissipated rapidly. Imagine that at a certain moment the expansion of a given volume of the medium ceased. Contraction should immediately follow because any rest is impossible once the force of proper gravitation is not balanced by any other forces. The unimpeded contraction under the effect of proper gravitation is essentially the free fall of the cloud s particles in their common gravitational field. However, free contraction cannot occur uniformly, at the same rate in every direction. However small the deviation from uniform contraction might be, which is unavoidable, it will be enhanced by the contraction itself. Indeed, if the rate of contraction in any of the three directions is by chance slightly greater than in the two other directions, the size of the cloud will decrease more rapidly in this specific direction, and therefore the compressive force of gravitation in this direction will become greater. This, in turn, will result in a greater contraction in the same direction and therefore in still greater flattening of the cloud. Thus an initially more or less spherical cloud is bound to become a flattened formation as a result of contraction.

Contraction leads to a gradual increase in the gas pressure, which does not play any role at first, but sooner or later should halt the contraction. Before that, however, a new phenomenon occurs that radically influences the further development of the cloud. When the gas fall speed in the cloud becomes comparable to the speed of thermal motion of its particles, or, which is the same, to the velocity of sound in the given medium, and later exceeds it, the gradual contraction of the gas is replaced by a so-called shock wave. This entails a drop in pressure, temperature, and density in the compressing gas.

A cosmic shock wave differs greatly on its scale from any "terrestrial" shock wave, but essentially this phenomenon is similar to the shock waves appearing, for instance, in an explosion or during the flight of a supersonic aircraft. In all cases, shock waves appear as supersonic motion of the gas, i.e. as the motion with a velocity exceeding that of the thermal motion of the medium's particles.

Since the role of shock waves in the cosmogony of galaxies and their clusters is essential, let us discuss them in more detail. This phenomenon belongs to nonlinear hydrodynamic processes, i.e. the processes in which perturbations of hydrodynamic values cannot be regarded as weak. Consider an example of a common sound wave.

This is a wave of perturbations of density, pressure, and velocity, where sites of maximum density alternate with sites of minimum density; they are distributed periodically and propagate in the medium with the velocity of sound. The distance between two neighbouring maxima of density is the wavelength. This is a weak perturbation where the relative value of the density amplitude (i.e. the greatest value of density divided by the average density of the medium) is small as compared with unity. And still, if we take a sound wave of large enough amplitude, the pressure and temperature in the maxima of density prove to be noticeably greater than their average values. The velocity of sound at these maxima is greater as well. This is why the crests of the waves propagate in the environment faster than the wave as a whole. Just as well, the velocity of sound in the minima of density is less than the average velocity, and therefore the troughs move somewhat more slowly than the whole wave; the result is that the crests tend, as it were, to overtake the troughs. When a crest gets closer to a trough, the layer of the density drop becomes narrower, or, as they say, the wave front becomes steeper (Fig. 12). Finally, there could be an instant when a crest could catch up with and even overtake a trough: the front would then turn over. Thus sea waves tumble over when they approach the shore, but these are waves on a free surface rather than those in a bulk. However, there is no real overturn in the kind of wave under consideration: when the distance

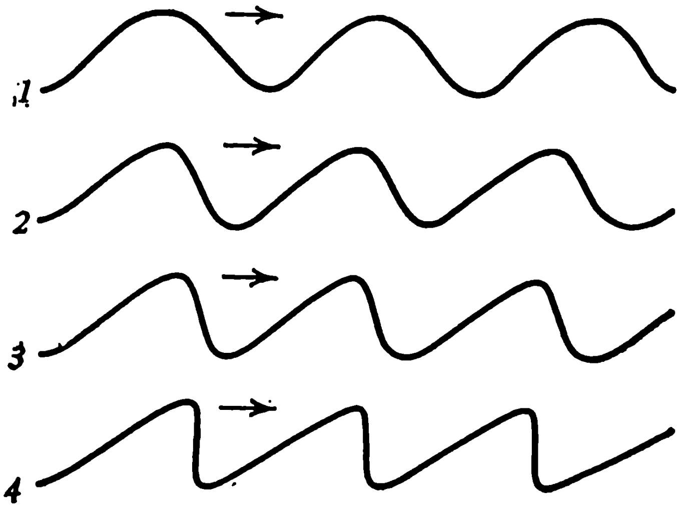

Fig. 12
The evolution of a sound wave. Density profiles are shown at four successive instants. The arrows show the direction of the wave propagation.

from a crest to a trough, i.e. the width of the wave front, becomes comparable to the length of the particles' free path, a shock wave appears, the whole front, i.e. the narrow layer where a sharp change in the hydrodynamic values occurs, propagating with a velocity greater than the velocity of sound in the medium.

Shock waves can appear both gradually (as has been described above) and abruptly, for instance, when a piston pushes gas through a cylinder with a supersonic velocity or when gas masses collide with such a velocity.

Let us return to the shock wave in a contracting cloud, i.e. a protocluster. It appears in a manner somewhat different than in the sound waves discussed above. At the beginning of contraction, a cloud is more or less uniform in its density; while it contracts and flattens, its inner layers become denser than its outer layers, and the cloud becomes increasingly nonuniform. The inner layers become heated in the process of contraction, and therefore the pressure in them increases owing to the thermal motion of the particles. The drop in pressure from the innermost layers to the outer layers gradually begins to decelerate the contraction of the inner layers. Therefore their free fall due only to the force of gravitation is replaced by a slower contraction. However, the outer layers continue their free fall, and therefore they gradually "press tighter" against the inner layers. The change in velocity (and the density with it) from layer to layer occurs when the layers get closer together. And the difference in the velocities of two approaching layers exceeds the velocity of sound in the medium. Sooner or later this results in a phenomenon similar to what happens with a sound wave: steep drops in density and velocity appear on both sides of the inner layer, i.e. two shock waves are produced separating the compressed gas of the inner layers from the outer "fresh" gas.

In the course of time, greater portions of gas in the condensation flow through the shock wave fronts to the area between the fronts, undergoing contraction and deceleration. Some of the kinetic energy of the gas flowing against the shock waves is transformed into heat at the fronts. Therefore the compressed gas is additionally heated and its temperature increases still more.

The structure of the hot and dense layer, i.e. a protocluster, was investigated by Ya. B. Zeldovich and his colleagues. They termed these gas layers "pancakes". Figure 13 presents the distribution of the concentration of particles (i.e. their number per unit volume) and the gas temperature in a "pancake" with due regard to the process of heating and cooling. Approximately half of the gas between the shock wave fronts belongs to the layer's hot fraction which is adjacent to its boundaries. Here the temperature amounts to tens of millions of kelvins; however, the gas density remains small, and the concentration of particles is about 10^{-3} cm^{-3}. The cooler inner layer contains gas with a temperature of about ten thousand kelvins, and the concentration is a thousand times greater there than in the hot gas. There is also a comparatively thin layer where the density is a few tens of times greater, while the temperature is respectively lower. The structure of the layer is such that the pressure in it is everywhere practically identical, so the product of temperature and concentration almost does not change from the central plane to the shock wave fronts.

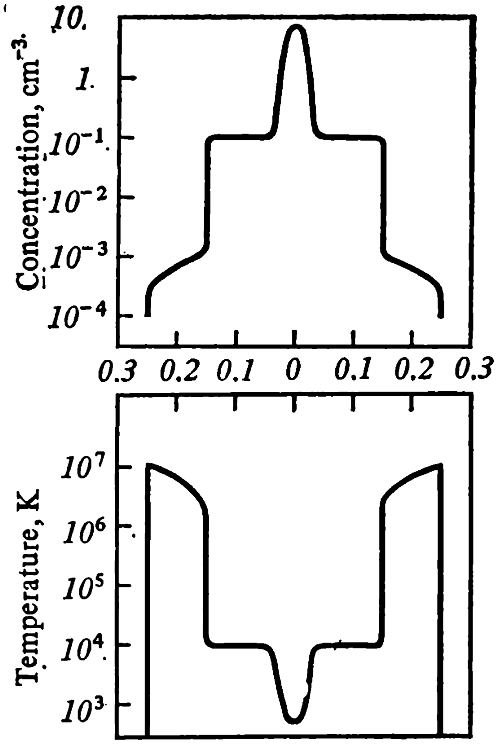

Fig. 13
The distribution of the concentration of particles and the gas temperature in a layer protocluster "pancake" at the moment when half of the mass of a cloud is between the shock wave fronts. The horizontal axis shows the fractions of the mass of the cloud on both sides from the central plane of the layer. The shock wave fronts are where the mass fraction is 1/4. Logarithmic scales along the vertical axes.

It is assumed that the dense inner areas of the layer, i.e. protocluster, undergo fragmentation and give rise to galaxies, while the hot gas fraction may be retained till the present. The proper gravitation of matter turns a cluster of galaxies thus produced into a gravitationally related and steady-state system. It is also assumed that the condensation gradually becomes more spherical and acquires the shape of a regular cluster like that of the Berenice's Hair (Coma) constellation.

Regular clusters exhibit a more or less pronounced concentration of galaxies about the centre. The predominant type of these clusters is elliptical galaxies. The Coma cluster and some other major clusters of galaxies produce X-ray radiation. It has been established that X-rays are emitted by the intergalactic gas in the clusters rather than by galaxies, the gas temperature reaching tens of millions of kelvins. The scheme of cluster formation presented gives a natural explanation to the origin of the hot intergalactic gas.

One of the central points in such a scheme is the disintegration of cool dense areas and the formation of galaxies out of them. So far this point has been in the stage of theoretical development, and it is not quite clear whether these areas disintegrate into protogalaxies or, perhaps, several stars or star groups appear first and then join together into galaxies. The complicated hydrodynamics of the forming clusters is being actively investigated by many astrophysicists, and the theory of "pancakes" is a solid basis for it.

The Large-Scale Structure of the Universe

The processes resulting in the appearance of clusters of galaxies developed against the background of a more colossal hydrodynamics, i.e. motions of a far greater scale, which shaped superclusters and possibly the largest formations in the universe, viz. the cell structures. Superclusters are groups of three to five clusters of various masses and sizes, which have been studied by astronomers during the past 30-35 years. The cells began to be discovered during the period of 1978-1982.

Both Soviet and American astronomers revealed that groups, clusters, and superclusters of galaxies are predominantly located in a comparatively thin layers or even chains. Such is, for instance, the chain of clusters of galaxies in the direction of the Perseus constellation. These layers and chains are connected to each other, they cross each other, etc., and form cells of irregular shape, the overall pattern being a quasi-ordered structure resembling lace or a honeycomb. The size of cells reaches a hundred megaparsecs. Their walls are superclusters of a very flattened shape (the ratio of the characteristic dimensions is about 1:5), and there are practically no individual galaxies or groups of them within the cells.

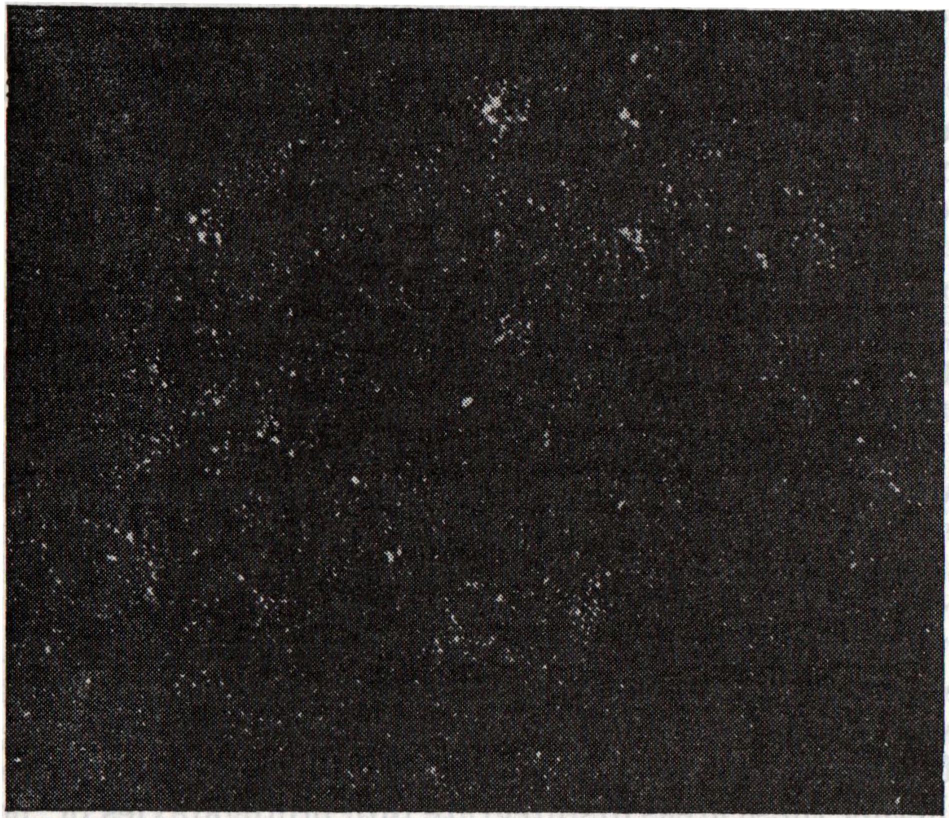

Fig. 14
A diagram of large-scale distribution of galaxies (according to J. Peebles).

However, it has to be said that these investigations are still in their infancy; there are reliable data on only three or four cells, i.e. voids surrounded by chains of clusters. There is no unanimity amongst astronomers in the estimation of the basic parameters of these systems, and sometimes doubts are cast with respect to the actual existence of the cell structure. In fact, so far there is insufficient observational material, and sometimes superclusters are revealed tentatively while considering the apparent distribution of galaxies over the celestial sphere. Galaxies look like points on the large-scale map of the universe, and these points are, in general, rather randomly scattered (Fig. 14). Superclusters exhibit themselves on such a map as fibers. However, it has long since been known that the human eye is apt to reveal linear sequences even in a completely chaotic distribution of points on a plane. 'To become convinced that the chains, cells, etc. are not illusions, some research is evidently required for a thorough selection of actual structures of the kind and their separation from illusory ones.

A detailed evolutionary theory capable of explaining the phenomenon of cell structure (assuming that it does exist) was suggested by Ya. B. Zeldovich and his colleagues A. G. Doroshkevich, S. F. Shandarin, and A. A. Klypin. Extending the theory of "pancakes", they also apply it to isolated clusters and whole superclusters inside the "walls" of the cell structure. The process of spherization in superclusters most likely occurs more slowly than in clusters of galaxies, and therefore they retain their original shape. The theory explains, therefore, why superclusters are so considerably flattened.

Moreover, it has been revealed that the "pancakes" are always shaped in an interaction with each other, rather than separately, and they share, as it were, the entire available matter. Enormous voids appear to be bounded by the "pancakes", or their walls, which join each other in something resembling a cell structure (Fig. 15). The theoretical model underlying this picture is mathematically complicated and is developed with the aid of computer-based investigations. Separate elements of the continuous medium are presented in the experiments as "heavy points" which possess masses and interact through mutual gravitation. A computer can calculate the motions of the whole set of these "particles", and this becomes a basis for the further consideration of the evolution of the medium from the initial state of weak perturbations to the state of strong perturbations.

The cell structure also appears in a quite novel version of the theory of "pancakes" taking into account the possibility of a nonzero rest mass of neutrinos. It is assumed that condensations of the greatest size corresponding to superclusters of galaxies are shaped first; the gas trapped by these condensations (now neutrino "pancakes") undergoes compression and heating, and then there occurs the fragmentation of its most dense layers into protoclusters and protogalaxies. Within the same theory, the main mass of clusters belongs to neutrinos, while the galaxies in them are some ten times lighter in compliance with the general cosmological relationship between neutrinos and other particles.

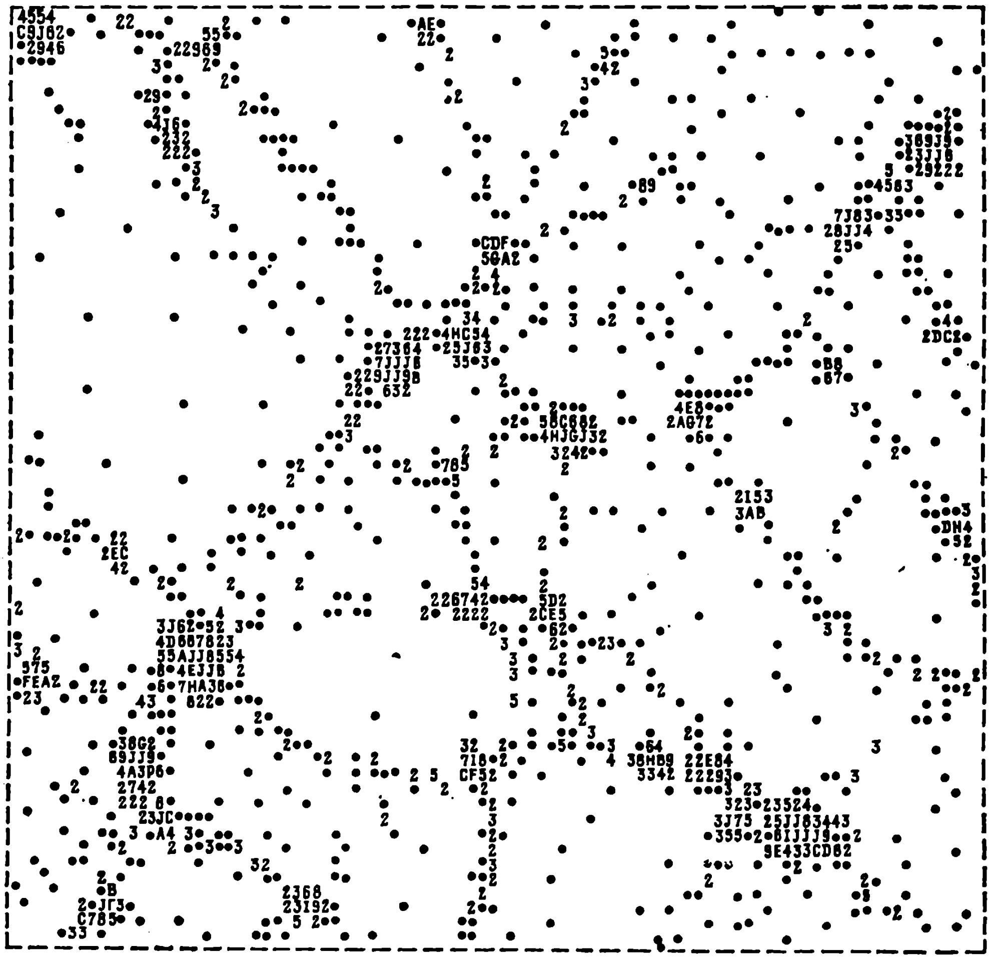

Fig. 15
The cell structure produced by computer-based investigations of gravitational instability.

Chapter 3

Stellar Eddies

The pictures taken using large telescopes show spiral galaxies as bright eddies of star clouds (Figs. 16-18). For instance, this is how the famous Andromeda Nebula (or the Andromeda galaxy) looks like, which is a gigantic spiral galaxy closest to us (Fig. 17). The luminous condensation in its centre branches out into wide and long spiral arcs sparkling with bright stars.

At great distances, our own Galaxy should look about the same. Its spiral pattern need only be derived through indirect data: clouds of gas and dust screen from us the stars of the galactic disk, and only the radio emission of neutral hydrogen, which is almost not absorbed by the clouds, indicates that the interstellar gas belongs to extended patchy and loose arms stretching from a distance of 3 kiloparsecs (kpc) from the centre of the Galaxy to its boundary, i.e. to distances of 15 kpc.

For an astronomer, a spiral pattern is an unmistakable indication of a galaxy's fast rotation. To be more accurate, flat subsystems of galaxies, like the disk of our Galaxy, rotate rapidly. A spherical subsystem, e.g. the halo of our Galaxy and the halos of others, rotate at least several times more slowly than their flat subsystems (disks). The rotation of the disks is, in a certain sense, maximum: the related inertia is such that it is exactly compensated for by the centripetal gravitational forces of the galaxy. If the rotation were faster, the stars of the disk would fly away from the galaxy owing to inertia.

Rotation is a very common property of galaxies. In act, all of them rotate more or less. However, elliptical galaxies do not rotate faster than the spherical subsystems of spiral galaxies. No essential rotation is noticeable in rregular galaxies. The disks of spiral galaxies rotate very apidly. And spiral galaxies prevail in the universe: hey exceed the galaxies of other types in both their umber and mass.

What is the nature of galactic rotation? Until recently, here were two hypotheses. According to C. von Weiz-

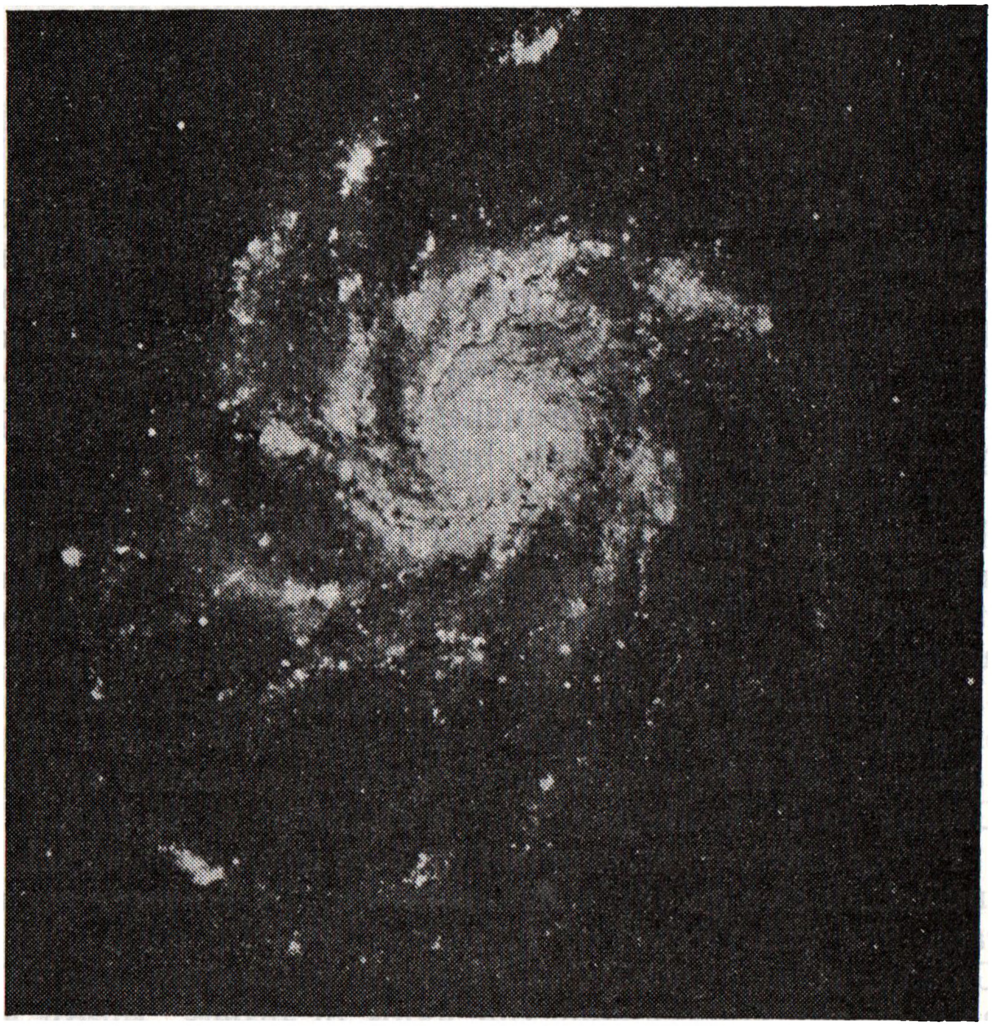

Fig. 16
A spiral galaxy in the Great Bear, or Big Dipper (Ursa Major) constellation.

säcker, who offered his hypothesis in the late 1940's, the rotation of galaxies is of cosmological origin; it is due to initial eddies generated in the cosmic medium during the creation and expansion of the universe.

During the same years, F. Hoyle offered another hypothesis, that of the tidal origin of galactic rotation. According to his idea, there was no rotation in the universe until the epoch of separation of protogalactic clouds. And it was only then that the tidal gravitational interaction between these clouds made them rotate.

The further development of cosmology and the physics of galaxies revealed the weak points and drawbacks

Fig. 17
The Andromeda Nebula, a gigantic spiral galaxy twice as massive as our Galaxy.

in both hypotheses. The first one, as revealed during the late 1960's, contradicts the concept of the hot universe. The second one, although it does not contradict anything,

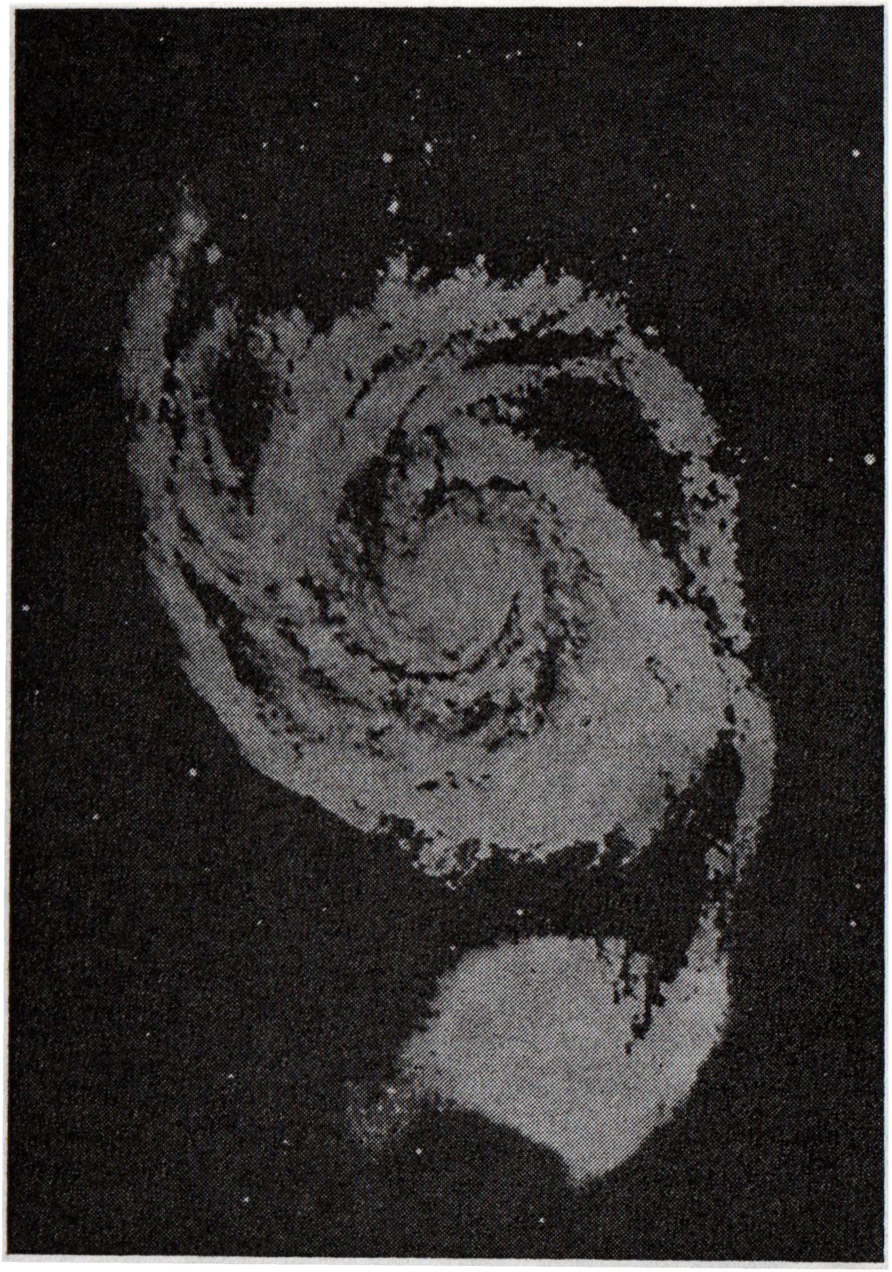

Fig. 18
A spiral galaxy in the Hunting Dogs (Canes Venatici) constellation.

cannot be confirmed by concrete calculations: it turned out that the tidal interaction between protogalaxies must have been too weak to produce the fast rotation of the spiral galaxies.

In 1970, A. D. Chernin suggested a new hypothesis on the nature of galactic rotation, according to which the rotation of galaxies is due to intense eddy motions appearing in discontinuous supersonic motions of the metagalactic medium.

This chapter deals with the first two hypotheses, which are both of historical interest and, what is of prime significance, related to profound and important physical ideas; we shall also treat in detail the third hypothesis, which is currently being developed on the basis of gravitational instability in the expanding hot universe. We shall give an account of the physical mechanisms underlying the appearance of the spiral pattern in rotating galaxies, too.

Rotation of Galaxies

The rotation of our Galaxy was discovered in 1926, when B. Lindblad and J. Oort established that stars in the disk of the Galaxy revolve about a common centre located in the direction of the Archer (Sagittarius) constellation. The linear velocity of rotation in the vicinity of the Sun (we revolve about the centre of the Galaxy together with it) is within 220 to 250 km/s; nobody has yet succeeded in refining this number. The rotation of other spiral galaxies has been measured directly in a few dozen cases; commonly the velocities of rotation amount to 100-300 km/s.

The linear velocity of rotation is different at different distances from the centre of a given galaxy. A typical dependence of the velocity on the distance (astronomers call it the rotation curve) is shown in Fig. 19. Till a certain distance, the velocity increases in proportion to the radius, and this implies that the corresponding inner area of the galaxy rotates with a constant angular velocity (as if it were a rigid body) and with a constant period. On reaching its maximum, the linear velocity gradually decreases to the boundary of the galaxy, and it proves that the dependence on the radius within this region corresponds to Kepler's third law: the velocity decreases in inverse proportion to the square root of the distance to the centre. The latter suggests that the revolution of the stars at the periphery occurs in the gravitational field created predominantly by the mass of the galactic central area, while the contribution of peripheral masses to this field is not very significant. The stars revolve at the boundary of a galaxy as the planets revolve in the gravitational field of the Sun (there are exceptions to this rule, see Chapter 6).

The period of rotation of a galaxy is assumed to be the period of its part rotating with a constant angular velocity, i.e. the period of rotation of the inner area. According to B. A. Vorontsov-Vel'yaminov, the periods of rotation of all observed spiral galaxies are within comparatively narrow limits, from 30 to 1000 million years.

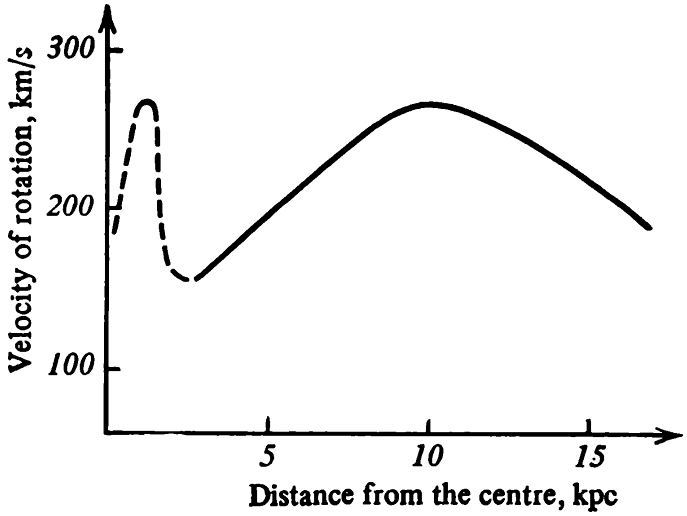

Fig. 19
A typical rotation curve of the disk of a gigantic spiral galaxy (the dotted line belongs to the central bulge).

The distribution of galaxies over their periods of rotation is uneven: there is a rather steep maximum around the value of 300 million years. Taking this value into account and assuming the typical age of a spiral galaxy to be 10,000 million years, we can find that spiral galaxies have only managed to make a few dozen rounds during their whole life span.

The rotation of galaxies can be characterized not only by the velocity and period, but by the angular momentum as well, also called the moment of momentum. Its value can be estimated as the product of the body's mass, its size, and the linear velocity of rotation. The angular momentum is a constant value for an isolated body (like energy, another conserved value). It can be calculated for our Galaxy assuming the velocity of rotation to be about 300 km/s, the size to be about 3×10^{22} cm, and the mass about 10^{44} g; the result is approximately 10^{74} g·cm^2/s.

The fact that the velocity of rotation of a galaxy at its periphery obeys one of Kepler's laws makes it possible to estimate the mass of the galaxy, or, more accurately, the mass of its inner area, which makes the greatest contribution to the force of gravitational attraction. Recall simple relationships that we all learn in high school.

Rotation with a linear velocity v at a distance R produces a centripetal acceleration of v^2/R. The acceleration of gravity produced by a mass M within the radius R is GM/R^2. The equality of both values yields the mass $M = v^2R/G$. If we take our Galaxy as an example and assume the velocity to be 200 km/s at the distance 10 kpc $= 3 \times 10^{22}$ cm from the centre (the position of the Sun), the mass is found to be about 10^{44} g, i.e. about 10^{11} Sun masses: the value commonly taken as the mass of our Galaxy.

Elliptical and irregular galaxies do not show any spiral structure. If their rotation is registered, it is always much slower than that in spiral galaxies. However, the slowly rotating galaxies are a minority among other galaxies: their number is three times less than that of spiral galaxies.

Eddy Cosmogony

In 1929, when J. Jeans studied the formation of galaxies during gravitational condensation, he assumed that galactic rotation is due to the rotational eddy motions which existed earlier in the rarefied medium from which protogalactic condensations were produced. So the spiral galaxies are, as it were, separated eddies.

Indeed, the shape of the spiral galaxies certainly indicates certain properties similar to those of eddies in air or water flows. Naturally, the prestar, pregalactic medium was not at all like, for instance, the Earth's atmosphere or the ocean, but the motions of very diverse media reveal some common properties and any example generally represents all other cases. The laws of mechanical motions of gas or liquid (fluid) masses are studied by hydrodynamics, a science which has existed since time immemorial, appearing during the observations of flowing water. To date, it has developed into a vast and comprehensive science of continuum mechanics, or classical field theory, retaining in its name, however, its first subject of investigation.

We have more than once discussed such hydrodynamic motions as adiabatic perturbations in the metagalactic medium, which were condensations and rarefactions in a mixture of plasma and radiation, and we have dealt with strong supersonic motions resulting from gravitational instability. Motions of this type do not contain any eddies: the adjacent layers of the medium can get closer to each other or more farther apart, but there is no relative rotation. Eddy motions are different. They can produce no condensations or rarefactions in the medium, and they are compatible with its complete uniformity. Besides an evident example of the general rotation of a gas mass (Fig. 20*a*), eddies can appear in such motions where, for instance, the flow velocity varies across the current (Fig. 20*b*). A straw in such a flow would not move parallel to itself but rather rotate.

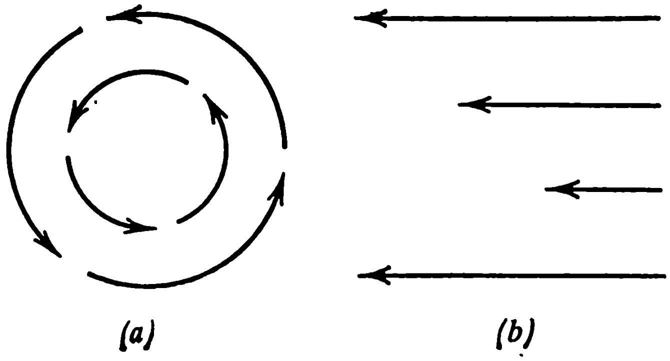

Fig. 20
Two examples of eddy motions. The velocities in the medium are indicated by the lengths and directions of the arrows.

A very essential feature of eddy motions is that they are "frozen" in the flow: an eddy involving some particles of the medium cannot be conveyed from them to other particles but is always exclusively related to them, and an eddy is displaced from one site to another only if the particles themselves are displaced.

The tradition of applying hydrodynamics to cosmogonical problems can be traced as far back as the science of the Renaissance, when its proponents used the images of cosmic eddies and other hydrodynamic ideas and notions in their hypotheses of the origin of the solar system. In 1609, J. Kepler was the first to draw the Sun in the centre of a powerful eddy pushing planets to their orbits and making them revolve about the Sun. Generalizing this hypothesis to the entire universe, R. Descartes wrote in 1644 that during the formation of cosmic bodies, the space of the universe was filled with an enormous number of eddies of diverse shapes and sizes. Newton did not leave these ideas unattended, although he regarded them critically and suspected that Descartes' cosmogony was incapable of explaining Kepler's laws describing the motions of the planets in their orbits. Newton criticized these ideas in his famous *Principia* (*Philosophiae Naturalis Principia Mathematica*, 1687), where he wrote, among other things, that the eddy theory neglected the observed astronomical phenomena and created more problems than it explained, making things more complicated rather than easier, etc. Newton wrote that there were no evidences of the eddies themselves, and therefore they should be rejected.

Later, I. Kant gave a profound analysis of problems of eddy cosmogony, and in 1796 P. Laplace deduced his nebular hypothesis from it, which has essentially been in the centre of cosmogonical discussions till now while being developed and enriched with new theoretical ideas and observational data. In 1911, H. Poincaré wrote, "Despite the numerous objections against it, despite the new amazing discoveries in astronomy surprising astronomers themselves, eddy cosmogony is still with us."

Protogalactic Turbulence

The ideas of cosmic eddies spread from the problem of the origin of the solar system to the new area of cosmogony, the theory of galaxy formation. Following J. Jeans, the hydrodynamics of pregalactic eddies was studied in 1948 by C. von Weizsäcker, who introduced to cosmogony the concept of hydrodynamic turbulence which had been developed by A. N. Kolmogorov shortly before.

Turbulence is a widely common natural phenomenon which always appears when motions of gas or liquid vary in time and space in a complicated chaotic manner. Turbulent motion with eddies and irregularities is contrasted in hydrodynamics to the smooth and regular motion which is called laminar motion. Turbulence has been studied for almost a hundred years, since the discovery of the apparently irreconcilable contradictions between theoretical hydrodynamics and experiments with the flows of gases and liquids. For instance, theory predicted that an increase in the velocity of a liquid flowing through a tube should bring about a resistance to the motion proportional to the velocity (according to Poiseuille's law). However, experiments showed that the resistance increases as the square of the velocity (according to Chézy's law) rather than its first power. The gist of the effect has become clearer since 1883, when O. Reynolds published his paper, reporting the results of his experiments with coloured water threads in a flow. O. Reynolds established a striking fact: the motion of the water threads was smooth and regular, they did not mix and remained distinctly separate from each other as long as the flow velocity was not too great. At higher velocities (all other conditions in the flow being equal), the water threads mixed and dissolved rather rapidly, colouring the entire flow more or less uniformly. When the threads mixed, it implied the development of a new phenomenon at the higher flow velocities: the smooth laminar flow became chaotic, or turbulent. Such turbulent "stalling" in a flow of water within a tube 1 cm in diameter at room temperature occurred at velocities exceeding 30 cm/s.

A swirling chaotic flow met a greater resistance than a smooth laminar flow, and therefore it was only natural to connect the results of the experiments with the turbulence. Theory could only calculate a laminar flow in the tube, and thus it was inapplicable to the given experimental conditions. This was the solution of a contradiction in hydrodynamics, which became the point of origin for the elaboration of a completely new concept in continuum mechanics, viz. the concept of turbulence.

The conditions of transition from laminar to turbulent flow were studied experimentally under diverse natural and laboratory conditions, and it proved that not only a greater value of the velocity was involved. While the velocities in identical tubes could be the same, some liquids exhibited laminar flow and others displayed turbulent flow; turbulence evolved earlier in liquids with lower viscosity, and very viscous liquids, such as honey, virtually could not flow in any mode other than laminar. O. Reynolds gave a criterion for the appearance of turbulence in the above-mentioned 1883 paper: the product of the velocity times the characteristic length of a flow (for instance, the radius of the tube) should be far greater than the coefficient of the kinematic viscosity of the medium. The dimensionless number equalling the density of a fluid, times its velocity, times a characteristic length and divided by the fluid viscosity is called the Reynolds number. More exactly, the Reynolds criterion is the principle that the type of fluid motion, that is, laminar flow or turbulent flow, in geometrically similar flow systems depends only on the Reynolds number; for example, in a tube, laminar flow exists at Reynolds numbers less than 2000, turbulent flow at numbers above about 3000.

While being random and apparently chaotic, the developed and steady turbulence can possess certain regularities. They cannot be revealed in separate eddies (vortices) where the velocity varies unpredictably from point to point and from one instant to another; the regularities are statistical and exhibit themselves in average characteristics of turbulent eddies. L. Richardson indicated in the 1920's that turbulence is composed of a set of eddies differing in their velocities and characteristic lengths. Eddies can interact, exchange energy, separate into motions of smaller scales or merge producing eddies of larger scales. But while isolated motions and their interactions are random, any set of eddies shows the same tendency to establish a kind of cascade of eddies, and the largest eddies (the largest in terms of both their size and kinetic energy) give rise to and feed the motion of eddies of smaller scales. When this tendency occurs completely, a universal relationship is established between the average velocity and the average size of an eddy in a turbulent cascade: the average velocity decreases along the cascade from top to bottom in proportion to the cube root of the size of the eddy. This property of developed turbulence was established by A. N. Kolmogorov in 1941.

Attempting to recreate the picture of the universe during the epoch of galaxy formation, C. von Weizsäcker surmised that the turbulence in the protogalactic medium encompassed masses of matter comparable to the masses of galaxies. It is easy to estimate the velocities and the characteristic lengths of the corresponding eddy motions.

If a galaxy has reached a typical density of about 10^{-24} g/cm^3 while contracting from the initial density of about 10^{-27} g/cm^3 characteristic for gas protoclusters (see Chapter 2), then the size in its initial state was apparently ten times greater than in its final state. It means that a galaxy such as ours had an initial size of about 3×10^{23} cm. This is the characteristic length of protogalactic eddies. The respective velocity can be found using. the law of conservation of momentum. As we have already mentioned, this implies that the product of a body's mass, times the velocity of rotation, times the size is constant. It follows that if the size decreases, the velocity of rotation increases in inverse proportion to the size. Such a gain in the velocity of rotation occurred during the contraction of protogalactic clouds. A ten times decrease in size means the same increase in velocity: if a developed galaxy rotates (at its periphery) with a velocity of, for instance, 300 km/s, this suggests that the eddy that generated it initially possessed a velocity of 30 km/s.

Eddies play a dual role in the cosmogonical hypothesis defined by C. von Weizsäcker: firstly, they gave rise to the rotation of the fragments, i.e. protogalaxies, and secondly, they promoted the isolation of fragments from the "continuous" medium (the continuum). It is as if the eddies were superimposed upon the general cosmological expansion so that each element of the medium participated in two motions at the same time: in the general expansion with Hubble's velocity and in the chaotic eddy motion with a random velocity. When the two velocities were comparable in value and antiparallel, then these motions in the volume were withdrawn from the cosmological expansion; but when the velocities were parallel, then they complemented each other, and this could very well happen in an adjacent volume. Thus condensations and rarefactions appeared; the condensations were areas where the general expansion was partially or completely suppressed, they were separated from the rest of the medium and could further compress under the effect of their own gravitation, therefore building up their density and turning into protogalaxies.

It is interesting that the condition of the comparability of the regular and random velocities is a direct indication of the age of the universe at the epoch of galaxy formation. Indeed, the regular Hubble's velocity is expressed using the age of the universe, i.e. the time from the beginning of the cosmological expansion to a given instant. The velocity of the cosmological recession of any two elements of the medium is just the distance between them divided by the age of the universe (within the accuracy of an inessential factor nearing unity). If we take the size of a protogalactic eddy as such a distance and consider Hubble's velocity to equal the velocity of rotation of this eddy, then it is easy to calculate that the age of the universe at the epoch of separation of protogalactic eddies amounted to approximately 3000 million years. The age of present-day galaxies is known to be 12,000-15,000 million years; then the age of the universe should now equal 15,000-18,000 million years. This estimate is in good agreement with all available data on the age of the universe.

It is noteworthy that the density of 10^{-27} g/cm^3 which we assumed for the protogalactic medium complies approximately with the value of the average density the universe should have had at the age of 3000 million years.

The metagalactic medium in Weizsäcker's hypothesis was supposed to be cold and nonionized, and it was regarded as a gas of hydrogen atoms with a temperature of no more than a few tens or hundreds of kelvins. Under these conditions the velocity of sound did not exceed several kilometres per second. But this implies that the velocities of the protogalactic eddies, viz. 30 km/s, were supersonic. Therefore C. von Weizsäcker discussed the supersonic turbulence in the pregalactic medium.

However, it has to be mentioned that since the 1940's-1950's, when this hypothesis was being developed, and till now it has not become quite clear what should be properly understood as supersonic turbulence. The Reynolds criterion is satisfied under the conditions in question: the viscosity of the gas at the above-mentioned temperature and density was such that the Reynolds number for the protogalactic eddies was much greater than the critical value. But the criterion itself, just as the entire theory of turbulence, were only elaborated for subsonic motions, and they can only be applied to these motions. Naturally, this was clear; nevertheless C. von Weizsäcker assumes that the regularities of developed turbulence could probably be applied to supersonic turbulence, and using these regularities, he gave estimates to the parameters of spiral galaxies.

Following A. N. Kolmogorov, C. von Weizsäcker regarded pregalactic turbulence as a set of eddies with steady average characteristics and a certain dependence of the average velocity on the eddy's characteristic length (the latter has already been mentioned). It was hoped that if this was true, then the statistical regularities in the distribution and motion of galaxies, i.e. the "frozen" eddies, could reflect the statistical properties of the turbulence which had produced them. There was strong observational evidence that the pregalactic medium contained eddies, namely, that the velocities of rotation of spiral galaxies were comparable to the velocities of their translational motion within clusters of galaxies. Translational velocities reflect the dynamics of the entire cluster: they make it possible to estimate, for instance, the mass of a cluster; rotational velocities indicate the development of eddy motions in the protogalactic medium. Therefore, there was a way to take into account and comprehensively apply the achievements of hydrodynamics so that (recall Newton s reproach) the observed astronomical phenomena would not at all be neglected.

However it may be called, turbulence or not, the hydrodynamics of chaotic supersonic motions is extremely complicated, and therefore no universal and consistent theory in this branch of science exists; each particular problem involves a great number of mathematical and conceptual difficulties. A detailed list of concrete problems whose solution could add up to a more or less consistent picture of "supersonic turbulence" was made by C. von Weizsäcker and his colleagues. However, despite a decade of intense effort with the participation of the best experts in hydromechanics, all of the tasks on this list have not so far been accomplished.

But this is not the end of the list of problems of cosmic eddies. The most fundamental problems are probably the very nature of eddy motions in the protogalactic medium and their origin in the expanding universe.

Primordial Eddies?

In 1952, G. A. Gamow, the founder of the hot universe theory and no doubt the most authoritative cosmologist of the period, decisively and enthusiastically supported eddy cosmogony. He said then that the distribution and motion of galaxies definitely showed the traces of pregalactic turbulence, that the physics of "supersonic turbulence" was uncommonly interesting and profound, and that he saw the most promising vistas there. As to the origin of pregalactic eddies, G. A. Gamow offered a thesis that eddies had existed in the universe from the very beginning and had the same nature as the cosmological expansion. Since the nature of the expansion remains, in effect, unknown, the problem of the origin of protogalactic eddies was therefore put on the list of the difficult questions in cosmology whose final solution could not be expected too soon.

However, if we assume that eddies accompanied the cosmological expansion "from the very beginning", then we have to make clear what they should have been during the earlier epochs of the expansion, prior to the formation of galaxies. And in this attempt an acute contradiction was immediately found. The point is that the velocities of eddies should have reduced with the cosmological expansion, and therefore if eddies possessing velocities of a few tens of kilometres per second were to have existed when the universe was several thousand million years old, a much faster rotation of the medium at the beginning of the expansion would have been required. Again, this is evident from the law of conservation of momentum as applied to an individual eddy.

As we have already mentioned, the order of the angular momentum, or the moment of momentum, can be estimated as the product of the mass involved in the motion, times the velocity of this motion, times its characteristic lenght. The "freezing" of an eddy in the medium (we have discussed this property of eddies above) implies that the total mass of particles involved in the eddy does not change with time. Therefore the values of the velocity and size should vary so that their product remain unchanged:

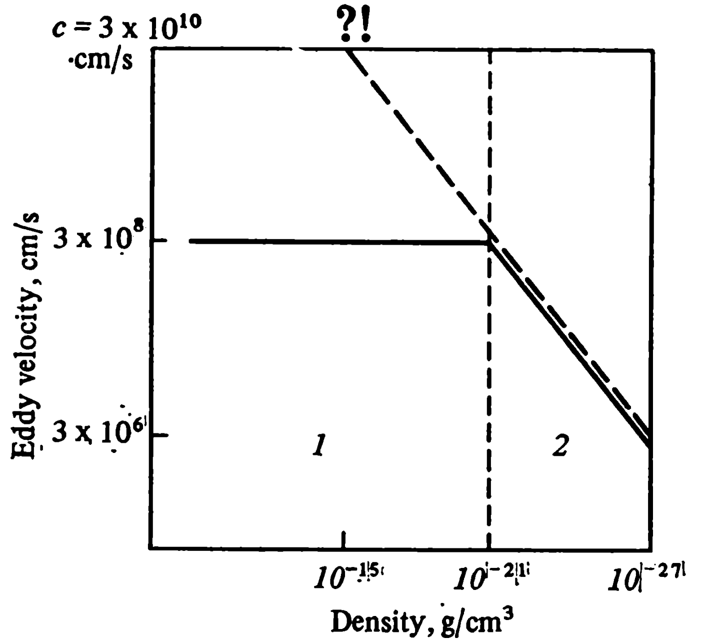

Fig. 21
Eddy velocity vs. density in an expanding cosmic medium. The dotted line represents the "naive" theory. *1*—the epoch of radiation prevalence; *2*—the epoch of matter prevalence. Logarithmic scales along both axes.

then the angular momentum would be constant as well. It follows that during the expansion the eddy velocity falls in inverse proportion to the size of a given eddy. The size of an eddy increases during the cosmological expansion in accordance with the general law which every length and distance obeys.

Proceeding from the dependence of the velocity on the size of an eddy, it is easy to establish its dependence on the density in the expanding medium: the velocity decreases with the expansion as the cube root of the density. This makes it possible to find the values of eddy velocities in the remote past of the universe, using the available data on protogalactic eddies.

During the epoch of separation of protogalaxies possessing a density of about 10^{-27} g/cm^3, the eddy velocities amounted to about 3×10^6 cm/s, as we have already mentioned. But it does follow that during an earlier epoch, when the density of the universe was 10^{12} times greater, the eddy velocity should have equalled the velocity of light, i.e. 3×10^{10} cm/s (Fig. 21). The density of the medium would still be very small at that (10^{-15} g/cm^3), and it would be very far, a good three hundred years, after the beginning of the expansion, i.e. the cosmological singularity to which we could refer anything.

At greater densities, eddy velocities would become superluminous, which is naturally absurd.

This paradox seemed to cast doubts on the entire field of Weizsäcker-Gamow eddy cosmogony. However, the contradiction was resolved after the discovery of relict radiation in 1965.

This discovery changed the concept of the properties of the metagalactic medium during the earlier epochs. Radiation, the photon gas, was the principal component of the metagalactic medium during the first million years after the beginning of expansion, and "common" matter was then less dense than photons. The hydrodynamics of the photon gas (with an admixture of plasma) differs from the hydrodynamics of a "common" medium which was implied in the discussion presented above. The behaviour of eddies in the photon gas is different, although momentum is naturally conserved.

The problem is that the mass of a given number of photons, into which an eddy was "frozen" through the plasma associated with them, did not remain constant, but rather declined with the expansion inversely proportional to the size of the volume occupied by the mass of photons. It follows from the conservation of momentum that the eddy velocity remained the same during the entire epoch of radiation prevalence. The velocity during the early epoch is easy to estimate taking into account that, starting from the moment when the densities of both components of the cosmic medium, i.e. matter and radiation, were equal (this occurred when they were about 3×10^{-22} g/cm^3), the "common" relationship between the density and the eddy velocity is valid, i.e. the velocity decreased in proportion to the cube root of the matter density during the later epoch. Then the eddy velocity at the initial moment when the densities of matter and radiation were equal amounted to 3×10^{8} cm/s, which is a hundred times less than the velocity of light (see Fig. 21). Therefore, the paradox of superluminous eddies in the hot universe is eliminated.

But the theory of primordial eddies revealed other problems, one of which was related to relict radiation. The eddy velocity we have just calculated was not too great, but eddies possessing such a velocity would have produced certain distortions, or fluctuations in the relict background: the photons "rotated" by the eddy motion of the medium during the early epoch should have been capable of preserving traces of this motion till now. Because of this the relict radiation energy flow received by radio telescopes should vary from site to site for the same reason we discussed earlier, in Chapter 2, in connection with adiabatic and entropy perturbations. The expected level of fluctuations for the eddy velocity of 3×10^8 cm/s is, according to calculations, such that they would be noticed while using the RATAN-600 radio telescope. However, no fluctuations of the kind have yet been detected by radio astronomers.

Another difficulty is of a purely cosmological nature. The uniform and isotropic universe described by the Friedmann model does not admit any primordial eddies in principle: it is either an isotropic universe without eddies or eddies without an isotropic universe. This conclusion comes from the extrapolation of the eddy state to the past of the expanding medium. We have already seen that this extrapolation, taking into account the role of radiation in the early universe, makes it possible to avoid the paradox of superluminous motions. There is yet another circumstance which is inescapably encountered in any attempt to extrapolate the eddies far into the past. The point is that, looking back into the past, we should see the primordial protogalactic eddy becoming more and more compact because its size would decline as we go back in time in proportion to every length and distance in the universe. But the distance to the horizon of events would decline still faster. Therefore, as it is evident from Fig. 22, there was such a moment in the past when the eddy was hardly within the horizon, and it should have been beyond the horizon at still earlier instants. And such an eddy should have unavoidably created distortions in the very structure of space-time, i.e. in the geometry of the isotropic universe. The farther the eddy was beyond the horizon, the more essential the distortions were; ultimately, while tending to ever greater values (ad infinitum, mathematically speaking) of the ratio of the size of the eddy to the distance to the horizon, the deviations from isotropy increased infinitely. This implies that the universe with primordial eddies just could not have been isotropic.

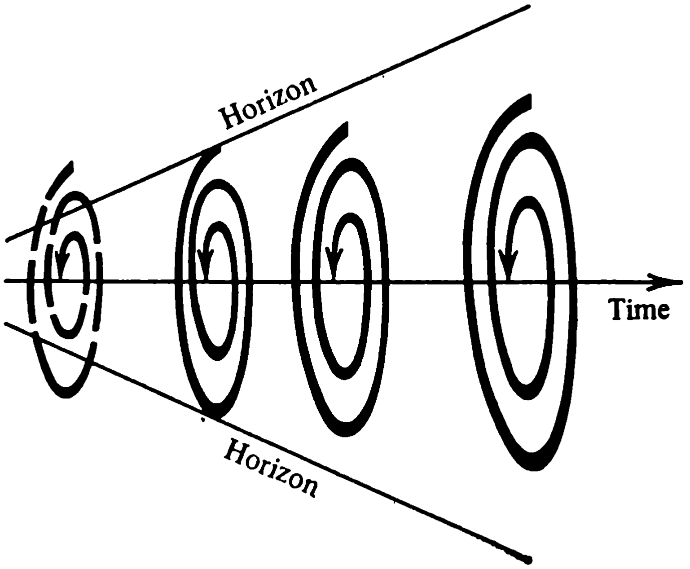

Fig. 22
Eddies and the horizon in the expanding universe.

Conceptually, the general theory of relativity admits nonisotropic (anisotropic) cosmological models, where no unbounded spatial symmetry underlying Friedmann's cosmology exists. In these models, the directions in space are not equal, so that, for instance, the rate of the cosmological expansion, i.e. the velocity of mutual recession of particles, is different in different directions. The expansion near the singularity in such models does not occur along all the three directions, and compression of particles may occur along one of them.

An in-depth study of anisotropic cosmological models was started in the USSR by A. L. Zelmanov in the 1950's. He found the conditions at which the initial anisotropy of the universe could be smoothed out with time so that the universe could exhibit the isotropy observed now. During the 1970's, E. M. Lifshits, I. M. Khalatnikov, and their colleagues revealed that quite unexpected behaviour of the universe was possible near the singularity: oscillations could be superimposed upon its general expansion. The oscillations alternately encompassed the motions along the three directions. The periods of oscillations declined as the beginning of the general expansion approached, while the total number of oscillations increased without limitation at the same time.

An analysis of the evolution of eddy motions in an anisotropic cosmological model shows that the eddies could be easily, as it were, integrated into the overall dynamics of the anisotropically expanding universe and could exist in it from the very beginning.

We have to mention, however, that no independent indications of an anisotropic beginning of the expansion have so far been found. On the contrary, everything indicates that during the early epochs, the universe as a whole was just as much isotropic as it is now. At any rate, there should have been no noticeable anisotropy during the recombination epoch when the universe was about one million years old. We can make a reliable conclusion on the isotropy of the universe during this epoch, considering the isotropy of relict radiation which was "torn" from matter at that time and freely propagated till now. Moreover, there are weighty theoretical deductions in favour of the original isotropy of the universe. The development of the concept of the quantum processes in the vicinity of singularity infers that if there had been strong deviations from isotropy "immediately" after the beginning of the expansion, then the effects of the creation of particles would have eliminated these deviations. The primordial eddies should have "died" along with the deviations, and everything would have been over very soon, by the epoch when the age of the universe did not exceed 10^{-41} second.

Finally, we should mention one more stumbling block in the hypothesis of primordial eddies. It is associated with the behaviour of eddy motions during the recombination epoch. The transformation of the ionized matter into neutral matter released particles from their connection with photons, and therefore the velocity of sound in the medium dropped rather drastically: from a value close to the velocity of light to several kilometres per second. The velocities of primordial eddies during this epoch should have been at least a hundred times greater than the latter value. Consequently, while being subsonic "from the very beginning", eddy motions became supersonic during recombination. But supersonic motions should have generated shock waves capable of compressing the gas. However, the density of the medium during the recombination epoch (about 3×10^{-22} g/cm^3) was already about a thousand times greater than that of galaxies; therefore the supersonic eddies of a galactic scale would have produced bodies that were too dense—nothing like the real stellar systems.

Adding up all these reasons, we should drop the idea of primordial eddies: most likely there were no eddy motions generated in the general process of the cosmological expansion in the universe.

Tidal Torques?

In 1969, looking for other concepts, J. Peebles analyzed from new positions the idea which was posed in the 1940's by F. Hoyle. According to this idea, the rotation of galaxies could have appeared without any eddies, just owing to the gravitational interaction between protogalactic clouds.

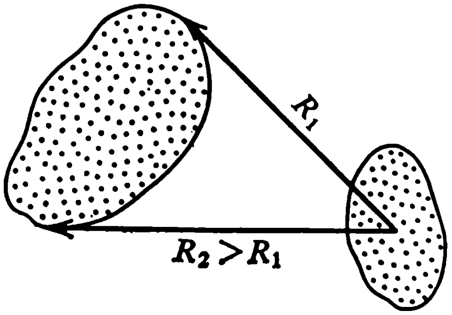

Fig. 23
Tidal interaction between protogalaxies.

This mechanism for the development of rotation requires that protogalaxies would be nonspherical and rather close to each other. The force of gravitational attraction between bodies is different in different parts of the bodies; because the bodies are nonspherical, the distances to different ends of a body from the centre of another body are different (Fig. 23). The force appearing as a difference in the gravitational forces acting on an extended body is said to be the tide-producing force causing a tidal torque. The tides on the Earth are generated precisely by such a force produced mainly by the Moon.

The difference of forces bring about a torque, or a rotational moment, which rotates the cloud. The influence is reciprocal, so that both clouds start rotating. Naturally, the conservation of momentum is valid in this case as in any other case: since the bodies interact, the conservation applies, rather than to the angular momentum of each individual body, to the total momentum of the pair of protogalaxies composed of their own momenta and the momenta associated with the orbital motion of the clouds with respect to each other. The protogalaxies may develop both parallel, i.e. identical, rotation and antiparallel, i.e. opposite, rotation.

The problem is whether this rotation would be as fast as the rotation we observe. Detailed calculations performed by J. Peebles and later by other theorists showed that the torque acquired by galaxies owing to their tidal interaction, even under the most favourable conditions, is five to ten times less than the actual torque of spiral galaxies.

It remains only to conclude that the tidal interaction between protogalaxies is ineffective, and it can be neither the only nor the principal factor giving rise to the fast rotation of galaxies.

The Birth of Eddies

The ideas offered by J. Jeans, C. von Weizsäcker, and G. A. Gamow on the eddies in the pregalactic medium and the obvious hydrodynamic similarities we have already mentioned are too impressive, attractive, and demonstrative to turn out to be totally invalid. But is it actually necessary to regard protogalactic eddies to be primordial?

Hydrodynamics has a strict Kelvin-Helmholtz theorem forbidding, under definitely specified conditions, the creation and disappearance of eddies. Both C. von Weizsäcker and G. A. Gamow tacitly assumed that the conditions of the Kelvin-Helmholtz theorem always held true in the universe. If so, the eddies existing in the cosmic medium during the epoch of galactic formation should have existed in it "from the very beginning": nothing else is possible.

The real picture of hydrodynamic motions in the expanding hot universe was far more rich and diverse than was supposed during the 1940's-1950's. The most striking phenomenon was the appearance of enormous shock waves compressing vast masses of matter into layers that would become protoclusters. Shock waves appeared when the universe was about 3000 million years old, and they were an unavoidable result of the evolution of gravitational instability encompassing the masses of matter comparable to the masses of the largest formations in the universe.

The appearance of powerful eddy motions during the same epoch was just as unavoidable.

The point is that one of the conditions of the conservation or nonappearance of eddies in the Kelvin-Helmholtz theorem is the absence of discontinuities in the hydrodynamic motions concerned. And shock waves are discontinuities in the velocity, density, and pressure of the medium.

Consider a simple example of the appearance of eddics in a shock wave. Suppose there is a parallel flow of gas meeting the front of a shock wave. Let the front be curved rather than flat, for instance, convex towards the flow, as shown in Fig. 24. The velocity of the flow at the shock wave front decreases in a jump; to be more accurate, there is a jump in the component of the velocity that is perpendicular to the front while the tangential component remains the same: this is the general property of shock waves. In strong shock waves, where the velocity of the flow meeting the front is far greater than the velocity] of sound in the gas, the component of the velocity that is perpendicular to the front decreases by a factor of four. The result is that the parallel and therefore eddy-free flow of gas becomes a divergent flow while it traverses the front. Imagine a "straw" thrown into such a flow: before meeting the front, it would drift in the flow parallel to itself, while behind the front it would also turn. This rotation is the indication of an eddy the flow acquired traversing the shock wave front.

The appearance of an eddy does not contradict the conservation of momentum because the total angular momentum of the flow remains zero as before. Prior to meeting the front, the angular momentum was zero; while passing over the front, individual elements gain an angular momentum, but in different parts of the flow (above and below the centre line of the flow in Fig. 24) the angular momentum of individual elements acquires opposite signs, so that the directions of rotation in the flow become opposite; the result is that the total of the angular momenta remains zero.

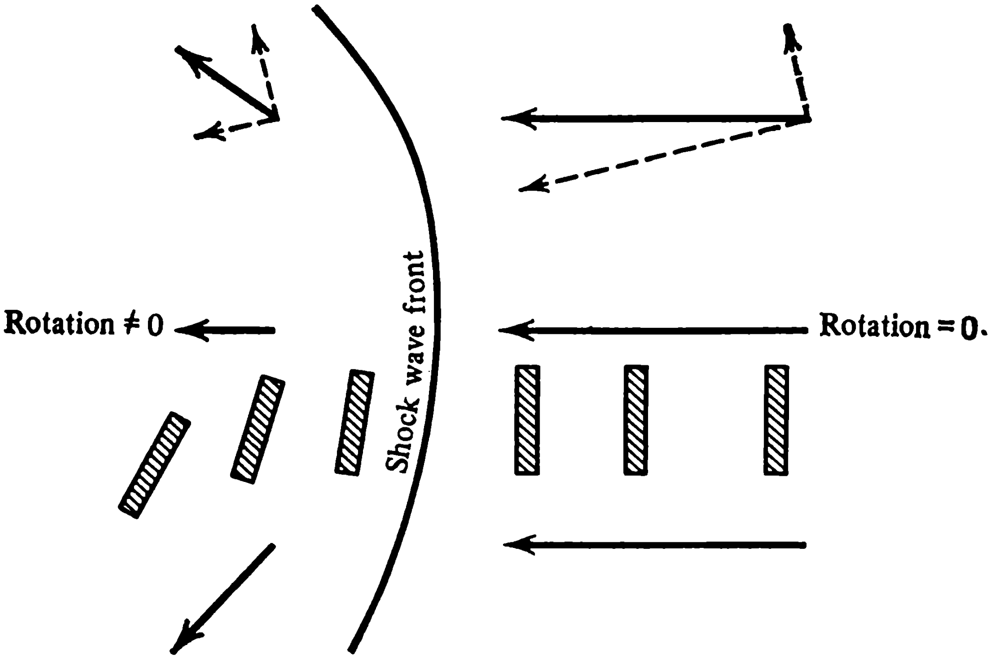

Fig. 24
The generation of an eddy in a flow crossing a shock wave front. The velocities of the gas are shown by the lengths and directions of arrows.

This mechanism of the appearance of eddies is only an example of a wide variety of processes which were able to develop effectively in the large-scale motions of the metagalactic medium and provide protogalactic condensations with a fast enough rotation.

Imagine a vast shock wave producing a massive protocluster, a "pancake". If the shock wave front is not curved but flat, eddies can appear on it all the same; this occurs when the gas meeting the front contains perturbations, i.e. condensations and rarefactions, or some comparatively small and loose clouds possessing proper motion. In contrast to our first example, the characteristic length of the appearing eddies is not commensurate with the characteristic length of the entire flow (i.e. its transverse dimension), but corresponds to the size of the initial clouds (perturbations). If the perturbations encompass masses comparable to the masses of individual galaxies, the condensations of such a scale, while passing over the front, become denser and then can turn into protogalaxies. The eddies which appear when condensations cross the front produce the rotation of protogalactic clouds.

The process of interaction between weak perturbations of a comparatively small scale (less than that of the whole motion producing the shock wave) has been studied in hydrodynamics for several decades. Initially, this problem appeared in an area having nothing in common with astrophysics, in the theory of flight of supersonic jet aircraft. Speeding up to the velocity exceeding the velocity of sound in air, the aircraft "pushes" and compresses the layers of gas close to its nose, and a shock wave whose front separates the compressed layers from outer noncompressed air spreads before it. This defines the resistance of air to the aircraft and, in effect, the entire dynamics of supersonic flight. It is essential to know how a shock wave is influenced by various heterogeneities in air and by perturbations in its density and pressure. Such perturbations are produced by the aircraft itself: its engine works and issues sound waves. These waves can reach the shock wave front because the front propagates with a subsonic velocity with respect to the compressed gas. But sound waves cannot travel beyond the shock wave front, the reason being that the front propagates in the nonperturbed gas with a velocity greater than the velocity of sound in this gas. Will the sound waves "get stuck" at the front, accumulate, and destroy it or will they, perhaps, reflect from the front—as a ball bounces from a wall—and travel backwards? It proves that reflection occurs, an echo appears, and the sound waves turn back without any damage to the shock wave.

The hydrodynamics of the pregalactic medium exhibits many features of this picture: there are strong perturbations in the medium, i.e. shock waves of the galaxy cluster scale, and weak condensations and rarefactions similar to sound waves, i.e. perturbations of the galaxy scale. The perturbations can either catch up with the shock wave front or meet it head-on. In the first case, the sound wave, as we have already mentioned, does not cross the front; it can be said that sound can reflect from the front without refraction (Fig. 25*a*). In the second case, the sound wave meeting the front travels beyond it and changes its direction of propagation, i.e. refraction occurs. There is no reflection in this case: sound cannot be reflected and "run away" from the front because the front itself travels in the gas faster than sound. In other words, the gas, in which sound propagates and meets the front, flows towards the front with a supersonic velocity and carries the sound beyond the front rather than allowing it to be reflected (Fig. 25*b*).

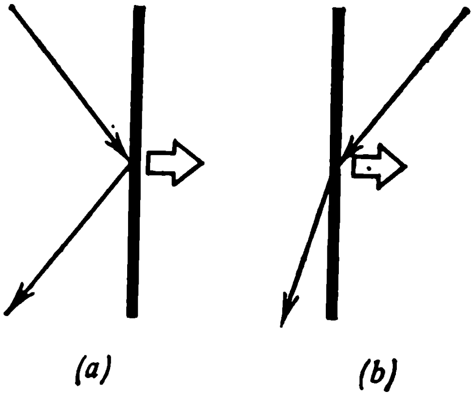

Fig. 25
A collision between a sound wave and shock wave front. The front moves through the gas from left to right, or, which is the same, the gas flows towards the front from right to left. The thin arrows show the direction of sound propagation: (*a*) reflection without refraction; (*b*) refraction without reflection.

Regarding these processes of interaction between perturbations and shock wave fronts, we are interested, rather than in the influence of perturbations on the shock wave front, in the transformation of eddy-free perturbations into eddies, which is a side effect in the theory of supersonic flight although it is a most important one in the theory of galaxy formation. There is only one case in which the perturbation remains eddy-free: when sound meets the front propagating at right angles to the plane of the shock wave front. Eddies are bound to appear in any oblique encounter of sound with a shock wave front.

It is interesting that when sound and a shock wave front travel towards each other, a phenomenon similar to optical total internal reflection can occur. When the light is incident on a boundary at a shallow enough angle, there may only be reflection without refraction, i.e. light does not cross the boundary between the media.

And in our case, when the angle between the direction of sound propagation and the perpendicular to the shock wave front is great enough (more than 60°), there is no refracted sound beyond the front. However, reflection is naturally prevented! But as we have already learned, eddies appear on the other side of the front, and they absorb almost completely the kinetic energy of the incident sound.

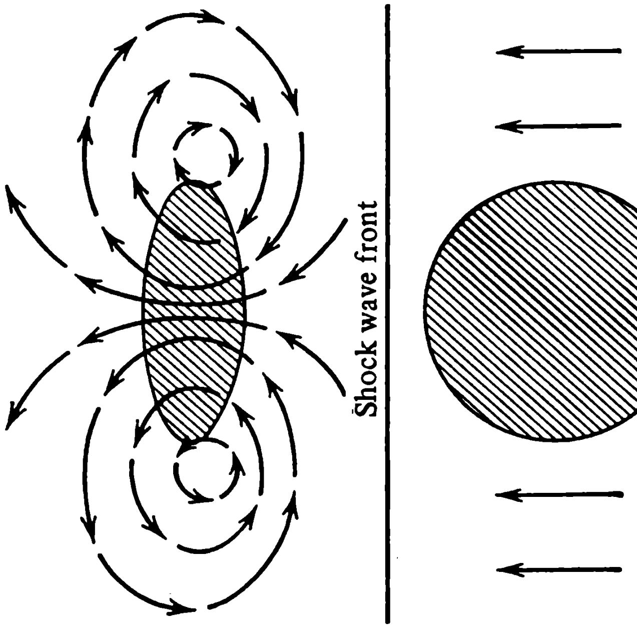

Fig. 26
The structure of eddies generating at a shock wave front while a condensation of matter passes through it (according to A.Yu. Ushakov and A. D. Chernin). The condensation area is hatched.

It is very essential that what has been said about sound is totally valid for any eddy-free perturbation. Any condensation of matter in a flow meeting a shock wave front generates eddies at the other side of the front (Fig. 26).

A Protocluster as a Turbulent Layer

The picture of supersonic motions in the metagalactic medium during the galaxy formation epoch is composed of a great number of diverse and complicated hydrodynamic processes. It includes both the formation of large-scale condensations (clouds) and the interaction between these clouds.

Let us take a step back and imagine pregalactic perturbations during the post-recombination epoch as a number of thin gaseous condensations, clouds, which were becoming ever more distinct owing to gravitational instability. For the same reason, each cloud as a whole acquired an ever greater proper velocity besides the regular velocity of the mutual cosmological recession of the clouds. While the proper velocities were small at the beginning, sooner or later they could have become comparable to the regular velocity of the expansion and even have exceeded it. Then the perturbations ceased to be weak, and hydrodynamic processes of a new type could have been expected to come into play.

Indeed, it is easy to imagine that if proper velocities were not small, then the mutual cosmological recession of two neighbouring clouds could be compensated for by their own relative motion which could, accidentally for this pair, draw these clouds closer together. Thus a collision of clouds was made possible. Evidently, loose clouds do not quite collide like elastic billiard balls. For instance, if two clouds have a head-on collision, they will not rebound but rather stick together and flatten. Their relative motion will be hindered by the collision of particles composing the clouds, and a layer of compressed gas will appear.

The collision of clouds will be nonelastic because the velocities at which they encounter each other are greater than the velocity of sound in the gas of each cloud. The motion of the medium cannot be smooth and continuous at such supersonic velocities. Jumps in density, velocity, and temperature are bound to appear in the medium.

These jumps can be of various types. One of them is a shock wave, where the component of velocity perpendicular to the front changes drastically, while the tangential component remains unchanged. As we have seen, if the flow meeting a shock wave front is eddy-free, eddies may appear in it when the front is crossed. Such phenomena invariably occur in the cosmic medium in nonelastic collisions of clouds created by gravitational instability.

Noncentral collision of gas masses is especially effective in producing eddies (Fig. 27). In this case, the material of each colliding cloud contracts and slides along the surface separating the clouds. This discontinuity of the tangential velocity cannot exist for long and disappears. The velocities of the initial sliding tangential motion give rise to eddy velocities in the layer of compressed gas.

The instability of a tangential discontinuity was apparent to Helmholtz, who noticed that the boundaries of the air threads from musical wind pipes get wound up as periodic spirals. This boundary is a tangential discontinuity separating the air that moves from air at rest. Another well-known phenomenon associated with the instability of a tangential discontinuity is the appearance of waves on a water surface under the effect of the wind. The discontinuity evidently appears on the surface of the water along which the wind blows: air layers move, and the water cannot stay calm because of this.

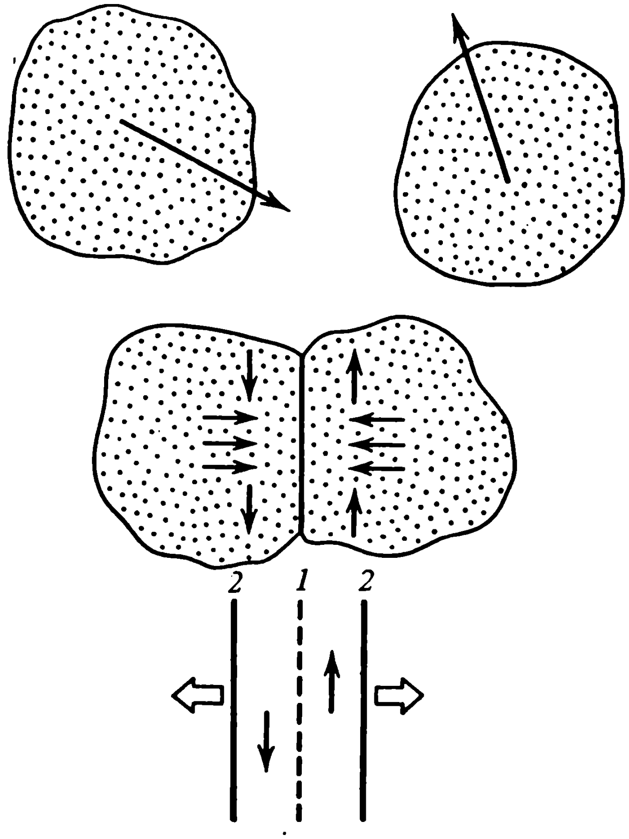

Fig. 27
A supersonic collision of gas masses. The generation of a tangential discontinuity (*1*) and two shock waves (*2*).

The nature of this hydrodynamic instability can be clarified by examining the behaviour of a weak perturbation, a distortion of the surface possessing a tangential discontinuity. If this perturbation is soon smoothed out and disappears, the surface of the discontinuity and the discontinuity as a whole are stable; but if the perturbation spontaneously develops and increases, the discontinuity is unstable.

A very simple case of a tangential discontinuity can be illustrated by the flow of a medium between two parallel planes (Fig. 28*a*). Suppose a fluid flows upwards near the right surface, while it moves downwards near the left surface (of course, no force of gravitation is taken into account here), and there is a jump in the velocity at the middle surface between the layers flowing in the opposite directions. Now suppose this discontinuity surface is distorted as shown in Fig. 28*b*. It can be said that the disturbance of this discontinuity surface has changed the cross-sections of the flows both left and right from the discontinuity: the left section became larger, while the right section became smaller. This is the reason why the flow velocities have to change at this site: the same quantities of fluid should pass through these altered sections. This means that the velocity on the left from the perturbation is less, and the velocity on the right is greater. Consequently, the pressure of the medium both left and right of the perturbation should alter. According to the classical theorem of hydrodynamics, the Bernoulli theorem, in a steady flow, i.e. such a flow in which the same quantities of fluid are transported through any section at any time, the smaller the velocity at a given site, the greater the pressure. The Bernoulli theorem, in effect, states that a wider section of a flow where the velocity is less than in a narrower neighbouring section requires greater pressure to produce the force needed to accelerate the portions of the fluid when they pass from the wid er

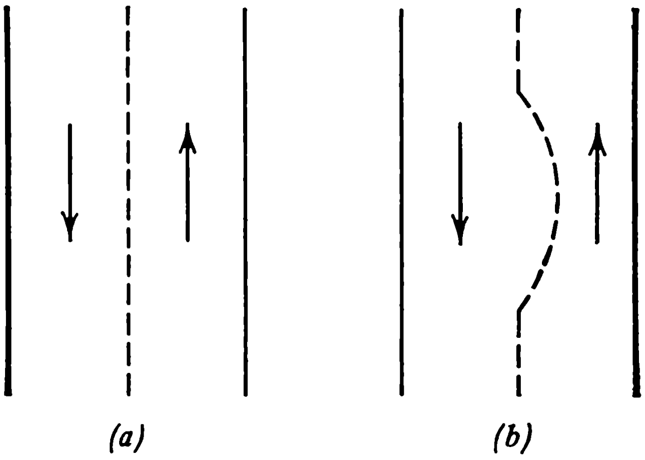

Fig. 28
The mechanism of instability of a tangential discontinuity.

section and therefore "push" them through the narrower section.

It is clear then that the pressure decreases before the convex part of the discontinuity surface and increases after passing it. This obviously makes the surface bend even greater. Greater changes in the cross-sections of the flows are therefore induced, the drop in pressure from left to right increases, and this results in the further development of the perturbation. This means that the tangential discontinuity is unstable.

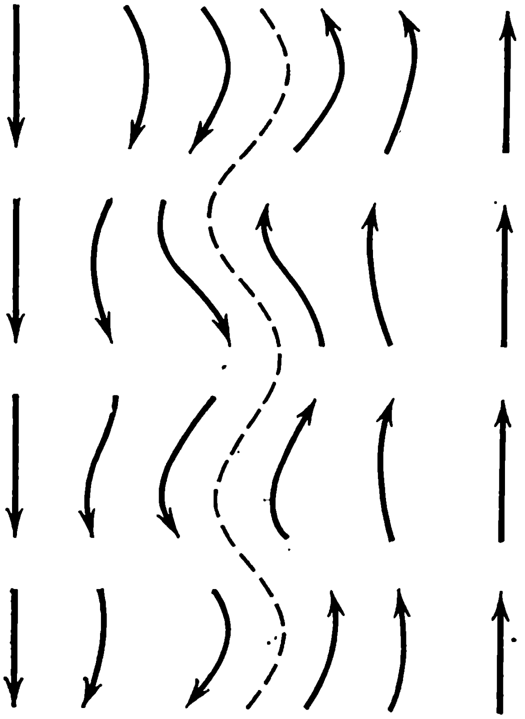

Fig. 29
The development of perturbations in a volume of a tangential discontinuity.

Naturally, the bounding surfaces in our example are not necessary. The point is the discontinuity itself, the jump in the tangential velocities between the neighbouring layers. Instability can exist in a free flow as well; it is due to the forces of pressure acting near the discontinuity surface.

The instability of a tangential discontinuity is not limited to a distortion of the discontinuity surface. It also causes perturbations in the neighbouring volumes of the medium. As seen in Fig. 29, the flow on both sides of the oscillating surface of a tangential discontinuity cannot be parallel, the velocities of the medium change, and these changes are the more intense, the greater the perturbations on the surface of the tangential discontinuity.

A tangential discontinuity strengthens any weak perturbations propagating through its surface. There is no flow of matter through this surface, but it can be crossed, for instance, by a sound wave. Amplifying a sound wave, the surface of the tangential discontinuity begins to oscillate, and this oscillation increases with time. The source of energy for the development of all these perturbations is tangential motions of the medium in the volume of a tangential discontinuity.

Returning to supersonic collision of clouds in the metagalactic medium, let us note that phenomena of this kind are an example of a strong and sudden external impact on a mass of gas. The appearing hydrodynamic discontinuities are said to be nonevolutionary because they are induced by an external cause rather than by the proper evolution of the motions. Nonevolutionary discontinuities appear, for instance, in a strong push of a piston compressing gas in a cylinder, in various explosions, etc. L. D. Landau pointed out that in every such phenomenon a whole system of hydrodynamic discontinuities rather than a single discontinuity develops. As far as our case is concerned, Landau's theory predicts that in the volume of initial contact between the clouds, a tangential discontinuity and two shock waves branching from it in the opposite directions should develop (see Fig. 27). The resulting layer of heated and dense gas is bounded by the spreading shock waves. The energy of the original countermotion of the clouds (perpendicular to the contact surface) is spent for the contraction and heating, while the energy of the sliding motion (tangential to the same surface) is stored in the tangential discontinuity. This contraction of a nonevolutionary origin resembles a "pancake"; by contrast, however, the "pancakes" discussed above are shaped in an evolutionary manner, i.e. owing to the gradual development of the gravitational instability in the given mass of gas. In both cases, there appear layers of heated and dense gas containing masses comparable to masses of galaxy clusters. However, a layer of a nonevolutionary origin differs from a "pancake" by the nature of internal motions: it is "charged" by the tangential discontinuity.

The development of instability in the vicinity of a tangential discontinuity leads to the appearance and development of various perturbations there, i.e. chaotic motions of various scales and intensities, drawing energy from the relative tangential motions of the gas. A characteristic feature of motions of this origin is always the presence of a considerable number of eddies in the general chaotic motion of the medium.

In fact, a tangential discontinuity is, as it were, an eddy in a plane (a vortex sheet). Such an eddy appears from an originally eddy-free motion of clouds when a supersonic collision of these clouds produces a nonevolutionary hydrodynamic discontinuity. It "has the right" to appear because if there are discontinuities, the Kelvin-Helmholtz theorem on the conservation of eddies, or the absence of eddies in our case, cannot be applied. The instability of a tangential discontinuity, as we have already mentioned, exhibits itself in that any weak perturbations present in the medium can be intensified in an interaction with the discontinuity. However, perturbations leaving a tangential discontinuity carry its energy with them but not eddies. When a perturbation interacting with a discontinuity is eddy-free, it becomes intensified but remains eddy-free.

However, there are no purely eddy-free perturbations in a protocluster. There are two reasons for this. Firstly, any perturbations crossing shock wave fronts together with a flow of gas acquire eddies at the fronts. This has been considered in detail above. Secondly, a perturbation cannot remain eddy-free even if it propagates within the compressed layer. The point is that the distribution of the density and temperature within this layer is not uniform; as in a "pancake", the density declines from the central plane outwards to the layer boundaries, while the temperature increases in the same direction. Under these conditions, the propagation of any eddy-free perturbations in a medium generates eddies. It is only necessary for the appearance of eddies that the disturbances should propagate nonparallel to the directions in which the density and temperature change. In a flat layer condensation, the density and temperature depend on the distance from the central plane and therefore change along the direction perpendicular to this plane. When, for instance, a sound wave propagates obliquely to this direction, the perturbation of its pressure depends not only on the perturbation of its density (as it happens in a uniform medium), but also on the behaviour of the "nonperturbed" density of the medium where the wave propagates. This kind of dependence (it is called nonbarotropy) makes the Kelvin-Helmholtz theorem on vorticity inapplicable; in this case, as in the case of a hydrodynamic discontinuity, this theorem is invalid, and eddies can appear and disappear, while purely eddy-free motion. is impossible.

The viscosity of the medium plays a very essential role in the development of instability of a tangential discontinuity. On the one hand, viscosity tends, as usual, to damp any eddy motions. On the other hand, it carries, as it were, eddies from the tangential discontinuity to the adjacent gas layers. Let us again recall the Kelvin-Helmholtz theorem and formulate its conditions in full: an eddy can neither appear nor disappear if a hydrodynamic motion has neither discontinuities nor nonbarotropy nor viscosity.

Viscosity is, in effect, the friction between the layers of fluid sliding along each other. When there is no viscosity (or, more accurately, when it is insignificant, i.e. it cannot noticeably influence the motion), the layers of fluid slide over each other freely, without interaction. If, moreover, there is a volume of flowing liquid with eddies and there are no eddies in neighbouring volumes (Fig. 30*a*), then the eddies will not penetrate into the neighbouring volumes (and naturally, the eddies will not be damped). But if there is a nonzero viscosity, an interaction occurs between the layers of the medium with and without eddies. The friction between the layers tends to eliminate the difference in the velocities; that is why the eddies in a viscous volume are damped, while they appear in the layers where there have been no eddies before. Indeed, while the velocities in these layers (on the left in Fig. 30*b*) were the same at the beginning, a difference between them appears because friction brings about deceleration, and this deceleration gradually propagates from right to left, i.e. from the volume where there are eddies to the volume where there have been no eddies before. Evidently, the decel ration produces eddies which penetrate into the adjacent layers and in due course fill the whole flow with eddies.

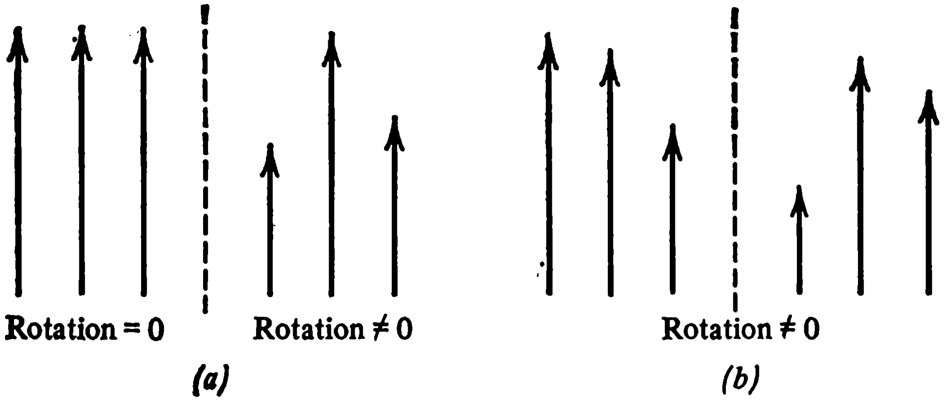

Fig. 30
The influence of viscosity: the deceleration and transfer of eddies between the layers of a medium; (*a*) nonviscous flow; (*b*) viscous flow.

While the instability of a tangential discontinuity develops, the pattern of gas current flowing along the surface separating the opposite flows of gas becomes ever more complicated (see Fig. 29). Generally, the more complicated the motion, the more significant the viscosity-becomes. It is insignificant and there is no friction between layers at all when every layer of the medium moves with an identical velocity (this is the simplest type of motion); but the friction between layers is unavoidable in a very complicated and nonuniform flow. The greater the difference in the velocities of the adjacent layers or the shorter the distances across the flow at which the velocities differ significantly, the greater the friction.

When the surface of a tangential discontinuity intensely oscillates and is distorted, the influence of viscosity exhibits itself in that it, as it were, switches over the fluid threads, or the streamlines, so that it produces isolated rotating volumes (Fig. 31). These volumes are called the cores of eddies (vortex cores). The axes of their rotation are parallel to the plane of the original tangential discontinuity; they are oriented like the eddies produced by the other processes we have considered above.

The appearance and enhancement of perturbations with significant vorticity and then the disintegration of a tangential discontinuity into vortex cores generate intense internal motions in a layer protocluster produced by a supersonic collision between some of the largest clouds in the metagalactic medium. The complicated and intricate nature of eddy motions resulting from the instability and disintegration of a tangential discontinuity gives rise to the development of turbulence in the gaseous protocluster. Thus the metagalactic medium acquires turbulent layers where gas is compressed, intensely heated, and what is especially important, possesses internal vorticity, i.e. rotational motions. While the total mass of the entire formation is comparable to the mass of a cluster of galaxies, the internal eddies encompass masses of gas comparable to masses of individual galaxies. The separation and gravitational condensation of eddies transform them into rapidly rotating spiral galaxies.

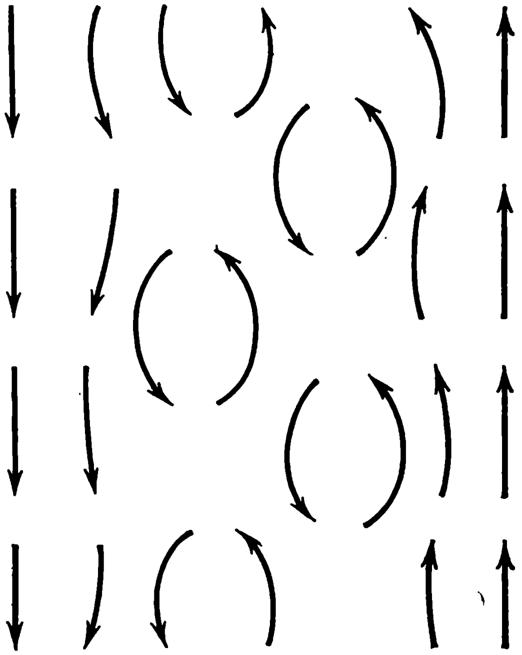

Fig. 31
The separation of vortex cores at the advanced stage in the development of instability of a tangential discontinuity.

Let us review the most important features of the hydrodynamics of a turbulent layer protocluster.

At the expected values of velocities and characteristic lengths of internal motions induced by the disintegration of tangential discontinuities, the Reynolds criterion indicating the appearance of turbulence is fully met for the physical conditions in a layer protocluster. Turbulence of this origin is characterized by considerable velocities;] however, they are less than the velocity of sound in the medium. The gas in the protocluster is strongly heated; as we have already mentioned, it was heated and made denser by the energy of the opposite motion of gas layers: this energy dissipates, or converts into heat, at the fronts of the shock waves bounding the layer protocluster. The temperature in typical protoclusters reaches tens of millions of kelvins, while the velocity of sound is several hundred kilometres per second. However, the velocities of the eddies capable of initiating the fast rotation of galaxies are dozens of times less.

The fact that protocluster turbulence develops under subsonic conditions makes it possible to apply here the general qualitative ideas and quantitative results elaborated in hydrodynamics.

The major eddies within the motions of a turbulent layer protocluster are vortex cores; their energy is capable of supporting smaller eddies, and it appears that as times goes by, there should evolve an orderly hierarchy of eddies with a cascade transmission of energy from a few major eddies to a larger number of smaller eddies. In the smallest eddies, the energy of hydrodynamic motions dissipates, i.e. transforms into the energy of thermal motions of particles under the effect of viscous friction. The turbulence of the Kolmogorov type seems to occur in medium-size eddies. Statistical isotropy is characteristic of the latter, i.e. the absence, on the average, of specific directions. Naturally, the major eddies are strong heterogeneities in the motion of the medium, and strong deviations from isotropy are associated with them; these deviations from isotropy do not disappear with averaging because the number of the major eddies is small: no more than, say, ten. But this heterogeneity is rapidly dissolved in a cascade of eddies, and the orienting influence of major eddies is not felt in eddies that are several times smaller.

Comparing the concept of a turbulent layer protocluster with the Weizsäcker-Gamow hypothesis, we can note that our concept completely retains the attractive features of turbulent cosmogony, along with its profound and imaginative physical content. However, a novel solution is offered for the nature of pregalactic turbulence: protogalactic eddies appear as regular effects caused by the whole preceding evolution of the metagalactic structure in the isotropic expanding universe. The same motions that produce large-scale condensations, or clouds, in the metagalactic medium give rise to the internal turbulent eddies in layer protoclusters. This explanation of their origin requires neither special assumptions nor hypotheses.

The concept of a turbulent layer protocluster helps explain the fast rotation of galaxies both qualitatively and quantitatively. Knowing the general characteristics of clusters of galaxies, we can estimate the velocities of the motions that produced these clusters. In order to halt the cosmological expansion of a mass of gas comparable to the mass of major clusters of galaxies (10^{15} Sun masses) when the universe was several thousand million years old, the proper velocities of individual volumes of the medium should have been close to a thousand kilometres per second. These are the velocities at which major gas clouds could have collided resulting in the formation of layer protoclusters. Because of the general chaotic motions of the clouds in the medium, the velocities of approaching and sliding motions were for the most part comparable to each other. And this means that the relative velocities in tangential discontinuities could also have reached a thousand kilometres per second. These velocities imply a large resource of kinetic energy, and this resource could have been spent (while tangential discontinuities disintegrated) for the initiation of turbulent eddies in the protoclusters. Naturally, some of this energy is lost, i.e. converts into heat owing to viscous friction. Therefore the velocities of the major eddies might be dozen times less than the original velocity of a tangential discontinuity. But this alone could have given rise to the fast rotation of galaxies appearing when the eddies separate. An eddy velocity of 30 km/s in the gaseous protocluster is sufficient for a galaxy such as ours to rotate with a velocity of 300 km/s (almost one radius from the centre).

We can assume that the clusters of galaxies, like the supercluster (in the direction of the Virgin (Virgo) constellation) to which our Galaxy belongs, were produced by turbulent layers. Most of the galaxies in these clusters are spiral, and there is a small number of gigantic spirals. Probably, the gigantic spirals are, in fact, separated vortex cores, while the spiral galaxies of smaller masses were produced by the eddies of a turbulent cascade. Clusters of this type are said to be irregular: they are patchy and do not exhibit an orderly general structure or visibly regular shape. This is, perhaps, a consequence of the violent and chaotic nature of the internal motions in the primordial turbulent layer.

By contrast, there is a smaller number of regular clusters of galaxies, like the one in the Berenice's Hair (Coma) constellation, of a regular spherical or elliptical shape. It can be thought that these clusters appeared from the protoclusters of evolutionary origin, or "pancakes". The internal dynamics of these protoclusters is calmer; there are no tangential discontinuities capable of generating an intense internal turbulence. It does not appear to be accidental that elliptical galaxies rather than spiral galaxies predominate in regular clusters, while the rotation of elliptical galaxies, if any, is much less powerful than that of spirals.

The Spiral Structure

Of the visible galaxies, no less than 70 per cent are spiral galaxies, and therefore it is only natural to assume that a spiral pattern is a frequent feature of a galaxy rather than a rare phenomenon. The high percentage of spiral galaxies also indicates that they are long-lived.

Most of the youngest and brightest stars are located in the spiral arms extending from galaxies. This is the reason why even the most remote galaxies have a distinct spiral pattern, although no more than several per cent of the galaxy's mass belong to the spiral arms.

Most frequently, there are two spiral arms wound in the same direction. There also occur more complicated spirals with three or four independent arms, and these sometimes branch out, too. The arms are always located in the plane of galaxy rotation. Most often, the arms are more or less wide open, but sometimes they are wound tightly and appear as rings at first sight. Some galaxies have a bar traversing the nucleus: if this is the case, the arms originate at the ends of the bar.

Figure 32 shows different types of galaxies in the tuning-fork diagram suggested by E. Hubble. The handle of the "tuning fork" represents elliptical galaxies, starting from perfectly spherical galaxies to lenticular galaxies, while the ends are spirals with or without a bar in the order of distinctness or the degree of opening of the spiral pattern. There was a time when astronomers thought this sequence was of an evolutionary nature, i.e. each galaxy was believed to have started as a spherical one, then to have become increasingly flattened, and finally, to have been developed by its spiral arms. This viewpoint is regarded as obsolete today. Modern concepts suggest that the type of a galaxy is determined by the conditions of its formation: if it was formed out of a rotating gaseous cloud, it had to become spiral, but if it was produced by a nonrotating (or weakly rotating) cloud, the galaxy would become elliptical.

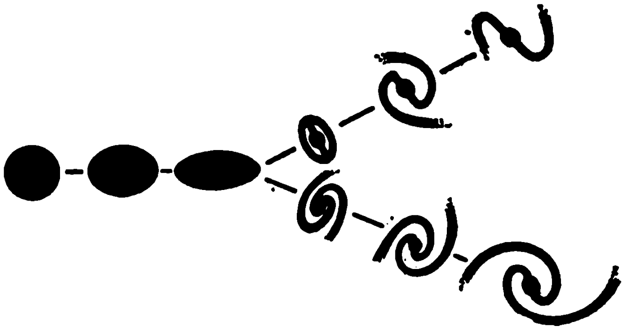

Fig. 32
Types of galaxies according to E. Hubble.

Then what is this spiral pattern?

When the study of galaxies was still in its infancy, it was surmised that the spiral arms are jets of luminous gas ejected from the centre of a galaxy and wound up into spirals by its own rotation. The gas produced stars that retained the original spiral shape of the gaseous arm by their location and motion.

Before long, however, it was discovered that spirals of such an origin would soon disintegrate because the rotation of galaxies is not uniform. Rotation is uniform, i.e. occurs at a constant angular velocity, as in a solid body, only in the inner area of a galaxy; but in the main bulk of the disk, i.e. beyond this area, rotation is differential rather than uniform, i.e. the angular velocity is not constant but changes (decreases) towards the rim of the galaxy. This nonuniform differential rotation could lead to the complete disappearance of the spiral pattern jets in three or four turns of the galaxy.

An additional factor is needed for a spiral arm to withstand the deterioration. Magnetic fields of galaxies were suggested as such a factor. It was believed that a spiral arm retains gas as a tube does, and the lines of the magnetic force lie along this "tube". However, it was discovered that the magnetic fields of galaxies are too weak to withstand differential rotation.

The modern concept of the galactic spiral pattern was formed during the past two decades on the basis of other ideas. In 1964, C. Lin and F. Shu suggested that the spiral structure be viewed as a wave propagating over the galactic disk. Similar ideas had been offered much earlier by L. Lindblad, but he had used rather complicated and unappealing mathematics. While the initial concepts compared the spiral pattern to jets or tubes containing the same particles from the beginning and during the entire period of their existence, the wave concept suggests that the spiral is a state of condensation propagating over the disk and passing from some particles to others. And in this case both particles of the interstellar medium and whole stars (the existing ones and those emerging in the galactic disk) are to be understood as particles.

A wave produces compression in the distribution of particles rather than drags them with it. It passes from some particles to others, producing compression out of new particles at a new site. This is a basic property of any wave process. The same happens to waves on a water surface: if a stone is thrown into the water, waves spread in circles, but water does not follow them. A straw would not be carried away by the waves but rather stay in one place and bob up and down.

A wave on the surface of water produces ring compressions if the water is not flowing. It looks quite differently if there is general rotation in the water: when it, say, drains down a funnel or it is stirred with a spoon as in making tea. Everybody knows that a wave in rotating water forms spirals rather than rings. And this is what happens in the disk of a spiral galaxy.

Naturally, the medium whose particles are whole stars is not very much like water. But wave processes in different media are very similar. The mathematical theory of waves in a rotating disk made of stars was developed by C. Lin and F. Shu and then refined by L. S. Marochnik, A. M. Fridman, and other investigators. It complies with the handy picture of spiral waves on the surface of water from our daily experience. Naturally, it takes into account the concrete properties of the galactic disk as the medium where the wave propagates. First of all, this theory considers that the disk is in its own gravitational field produced by its stars and the interstellar gas; the calculations involve the data on the distribution of densities and velocities in the disk, etc.

The main feature of a wave's spiral pattern is its uniform rotation. Although the disk of a galaxy rotates differentially, the spiral pattern rotates as an entity with a constant angular velocity. This remarkable feature eliminates every obstacle related to differential rotation that challenged the earlier theories. This is the reason why the spiral pattern is so clear and distinct and retains its regularity, undisturbed by differential rotation, over the entire disk of a galaxy.

Consequently, a wave produces compressions in the distribution of stars, and the wave crests appear as spiral arms extending over the whole galactic disk. But the density of stars in these crests is not so great. The star population is only a little denser there than in the disk on the average: there is no more than a ten per cent rise in the density. Such a poor contrast would never have been noticed in the photographs of distant galaxies if the stars in the spiral arms were the same as in the rest of the disk.

The point is that the interstellar gas gets condensed in the spiral arms together with stars, and it produces new stars there. They are very bright at the initial stage of their evolution and therefore stand out among the other stars of the disk. Observations of the neutral hydrogen in the disk of our Galaxy (by its emission at the 21-cm wavelength) indicate that the gas does produce spiral arms (Fig. 33). For young stars to outline the arms distinctly, a high rate of transformation of the gas into stars is needed, and furthermore, the duration of a star's evolution at its initial bright stage should not be too long. Both prerequisites seem to be met under the real physical conditions of the galaxies. The duration of the initial phase of evolution of bright massive stars is less than the time required for an arm to shift noticeably in its general rotation.

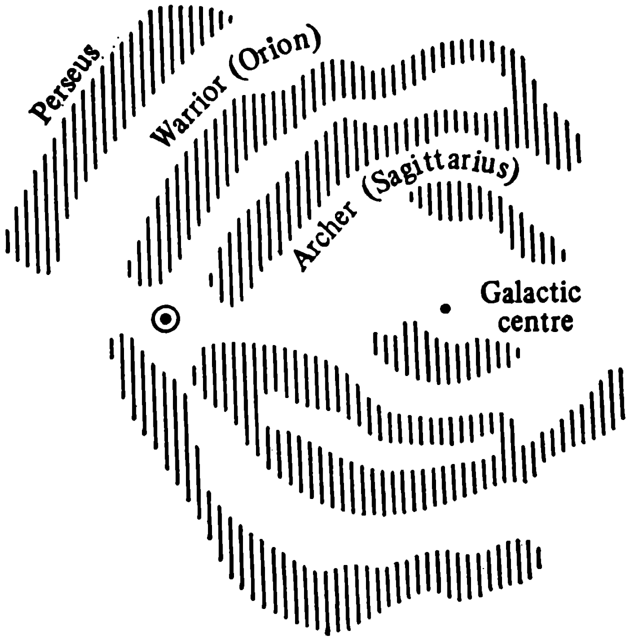

Fig. 33
A diagram of spiral arms of the Galaxy according to radio wave data on the distribution of atomic hydrogen.

W. Roberts and S. B. Pikelner developed an attractive hypothesis of star formation in spiral arms. The hypothesis is based on an analysis of the hydrodynamic processes in the interstellar gas caused by the relative motion of the gas and of a spiral arm. The velocity of rotation of the spiral pattern and the velocity of rotation of the galactic disk are different. Since this is so, the gas flows through the spiral arm, or, which is the same, the spiral arm moves through the gas. This relative motion is characterized by velocities greater than the velocity of sound in the gas. The gas can be regarded as moving through the arm with a supersonic speed. This causes a shock wave (again!) when the gas "collides" with the arm. The gas is strongly compressed in the shock wave, and this is, probably, the "triggering mechanism" for the process of star formation.

If the relative rotation of the gas and the spiral arm is the only cause of star formation, then one can expect that the star formation is hindered in the circle of the galactic disk where both velocities of rotation are the same, and the circle itself (the corotation circle) should be scarce in young stars. However, no dark rings in the spiral pattern have so far been observed. This probably means that, apart from the Roberts-Pikelner mechanism, there are other processes initiating star formation which do not require relative motions and for which even a small compression of the interstellar gas by a spiral wave is sufficient.

There are many difficult problems waiting to be solved in the theory of spiral structure, despite all its achievements. The most fundamental problem is that of the direction of the spiral wave: whether wave propagates from the centre outwards or, vice versa, from the rim of the galaxy to its centre. In actual fact, this is the problem of the "generator" of the spiral wave, i.e. the problem of where and how it is produced.

The mathematical theory of the spiral structure in a rotating gravitating medium demonstrates that any weak perturbation should propagate in such a medium as a spiral wave with a certain number of arms and velocity of rotation. But the observed spiral wave (which forms out of a variety of possible ones) with a velocity and quite a definite number of arms is probably controlled by the specific conditions of the wave generation in each given galaxy. C. Lin and F. Shu suggest that the generator of a spiral wave is at the periphery of the galactic disk and is a considerable condensation or, probably, a small satellite galaxy. The gravitation of such an object is capable of creating certain perturbations in the common gravitational field of the galactic disk, and these perturbations should generate a spiral wave in the disk. The wave of this origin travels towards the centre of the galaxy. The velocity of rotation of the spiral pattern is controlled by the velocity of revolution of the "generator" around the galactic centre, while the generator itself, be it a condensation or a satellite galaxy, should be found at the tip of a spiral arm.

There are quite a few galaxies possessing condensations at the tip of a spiral arm, e.g. the spiral galaxy in the Hunting Dogs (Canes Venatici) constellation (see Fig. 18). The external generator in our Galaxy may be located at 15 kpc from its centre, at the periphery of the galactic disk. The linear velocity of rotation there amounts to about 160-200 km/s. Evidently, this determines the angular velocity of the whole spiral pattern, which is the same over the whole galactic disk. It is usually expressed in km/(s·kpc). The linear velocity mentioned above corresponds to the angular velocity of 11-13 km/(s·kpc). The angular velocity of rotation of the galactic disk in the region of the Sun is about half as much, i.e. 20-25 km/(s·kpc).

The spiral pattern possessing two arms and rotating with this angular velocity does not seem to contradict the data on the distribution of the neutral hydrogen in the disk of our Galaxy.

However, these data are still not definite: they allow for an alternative. It is quite logical to assume that the "generator" is at the centre of the Galaxy rather than at its periphery. The source of spiral wave generation could be the hydrodynamic instability produced in the central part of the galactic disk owing to the specific properties of the rotation curve in this area. This idea was suggested by A. M. Fridman and V. L. Polyachenko.

Moreover, L. S. Marochnik and A. A. Suchkov suggest that a spiral wave can be generated by an asymmetrical entity whose rotation causes perturbations in the gravitational field of the galactic disk. This body should be essentially nonspherical because the gravitational field of a sphere does not depend on whether or not the sphere rotates. This may be the galactic bar, similar to the one observed in some galaxies (see Hubble's diagram, Fig. 32). The bar should rotate about its short axis in the middle.

The velocities of rotation in the central part of the Galaxy are greater than at its periphery. This is the reason why the spiral pattern of this origin rotates faster than that in the concept of an external generator. Most probably, the angular velocity of rotation is 20-

25 km/(s·kpc). And this is in keeping both with the data on the distribution of neutral hydrogen and the available data on the rotation of the inner areas of the Galaxy.

It is interesting that in this case the Sun should be located in the corotation circle, i.e. where the velocity of rotation of the disk of the Galaxy and the velocity of rotation of the spiral pattern are close to each other. L. S. Marochnik suggests that this is probably the reason why there appeared the necessary prerequisites for the formation of planets around the Sun and for the very emergence of life on the Earth. If this is so, it is worthwhile to search for other planetary systems and, who knows, for other civilizations precisely in the direction of the corotation circle in our Galaxy....

Chapter 4

The Birth and Evolution of Stars

Over nine tenths of the matter in our Galaxy is made up of stars, and there are galaxies where stars account for 99.9 per cent of the mass. The world of stars is diverse, but most of them are like our Sun. This chapter deals with the Sun and other "common" stars; we shall also discuss pulsars and bursters (which are quite unlike the Sun), the formation of the first stars in our Galaxy, the evolution of a star, the processes of present-day star formation, and finally, black holes, the terminal state of a massive star.

The Sun and Stars

The Sun, like other similar stars, is a spherical mass of hot gas kept intact by its own gravitation. The gravitation compresses the gas and draws its particles as close together as possible. The pressure developed by the hot gas acts in the opposite direction and tends to expand the gas. The force of gravitation is directed to the centre of a star, while the force of pressure is directed outwards; their counteraction establishes and maintains the equilibrium in which a star can survive for millions and thousands of millions of years.

The pressure in the Sun reaches 10^{10} atmospheres, and the temperature is estimated at 14 million kelvins. The great pressure and the high temperature are maintained in the central zone owing to continuous nuclear reactions of hydrogen transforming into helium. The released energy escapes as a flow of photons through the mass of the Sun to its surface and then outwards. The Sun emits 4×10^{26} J of energy per second. Most of this energy is carried out by photons (or, in terms of waves, by electromagnetic waves) in the visible part of the spectrum.

Taking into account the present-day rate of production and emission of energy, the Sun's "nuclear reactor" is well provided with fuel for more than 10^{10} years. This estimate proceeds from the fact that the transformation of hydrogen into helium releases (as emission energy) about one per cent of the rest energy of matter, i.e. about 10^{12} J per gram of hydrogen. Since the mass of hydrogen in the Sun is of the order of 10^{33} g, the resource of energy is about 10^{45} J. Dividing this value by the Sun's luminosity, we obtain the time for which the Sun is capable of emitting light owing to the nuclear reactions occurring within it.

The remarkable idea of nuclear transformations being the source of the luminosity of the Sun and stars was suggested in the 1920's by A. Eddington, the founder of the theory of star structure and evolution. The development of the physical concepts of star energy production on the basis of current nuclear physics and quantum mechanics was started in the 1930's by H. Bethe and then continued by many researchers throughout the world. This work, once commenced, is still developing today with due regard to the latest achievements in physics and astronomy.

Not every star but most of them are like the Sun; however, every star begins to exist as a sphere being at equilibrium owing to the balance of the forces of gravitation and pressure and heated from within by high-temperature nuclear reactions.

The life span of a star in such an initial state depends on the nuclear energy resource and the rate of its consumption. The greater the mass of a star, the greater its energy resource; however, the greater the luminosity of a star, the greater its mass. Considering large stars, three and more times more massive than the Sun, we find that their luminosity is proportional to the cube of their mass, which is known from both direct astronomical observations and the modern theory of the inner structure of a star. If a star, for example, possesses 50 Sun masses, its nuclear fuel can be consumed in several million years.

One can assume that the Sun is going to stay in its initial state for a long time, while the evolution of more massive stars occurs much faster, and when an essential part of the nuclear fuel is spent, significant changes of these massive stars should naturally develop in the structure and they should proceed far beyond their initial state.

Gravitational Condensation

Stars are formed and reach a state similar to that of the Sun owing to the gravitational condensation of rarefied clouds of gas. This idea goes back to Newton (recall his famous "cosmogonical letter" quoted in Chapter 2) and is the starting point for modern star cosmogony. As Newton pictured it, the fragmentation of the uniform medium and star formation occur owing to gravitational forces alone provided no other forces, if any, prevent it. It is interesting that Newton compared two cases, i.e. a finite and an infinite space (more accurately, volume) filled with matter. The matter in the first case should collect into one body, while its fragmentation into many separate condensations should occur in the second case. If the volume is infinite and matter is distributed in it uniformly, every point has equal rights, and therefore there is no special point to which "all matter in the universe" could gravitate as to the centre. There is no such single centre, and the uniform distribution of matter is unstable and cannot be at rest. Consequently, there should be an infinite number of centres around which separate masses of matter should collect.

But what is this infinity of space or volume in these considerations? In fact, infinity here means that the space size of the distribution is much greater than some other size used as a characteristic length, a yardstick. At the same time finiteness here means simply that these two sizes are commensurate, or close to each other. If the size of the area filled with matter is much greater than a characteristic length, the matter will divide into fragments; but if there is no such a strong inequality, the mass will remain compact and contract as a whole.

Most probably, this is what Newton implied in his cosmogonical letter of 1692; however, the nature of the characteristic length was only revealed by J. Jeans in 1902. Essentially, the space scale is determined by the forces opposing gravitation, such as the forces of pressure due to the elasticity of matter. This characteristic length has been mentioned already as the Jeans length.

In fact, the counteraction of the forces of gravitation and pressure in a rarefied medium has the same nature as in a star. But in a star these forces reach a "compromise", becoming equal to each other and therefore establishing an equilibrium. There is no such balance in a rarefied gravitating medium: as far as the medium is uniform, the force of gravitation reigns supreme. The force of pressure only appears when there are nonuniformities in pressure, i.e. drops in pressure from site to site. The force of pressure is always directed from an area of high pressure to an area of low pressure. Such drops in pressure appear when there are spontaneous condensations in the medium. The fate of a condensation and the future of the entire distribution of matter with it depend on which force turns out to dominate: the force of proper gravitation compressing the condensation or the counteracting force of pressure due to the drop in pressure between the condensation and the environment.

The greater the force of gravitation, the greater the mass and size of the condensation. Therefore it is capable of counteracting the opposing force of pressure if this size exceeds a certain critical dimension corresponding to the equality of both conflicting forces. The critical dimension is the Jeans length. And gravitational instability (the Jeans criterion) develops when the size of a condensation is greater than this critical length. One can safely say that the Sun and any other star where the forces of gravitation and pressure balance each other have a size equalling the Jeans length.

But the Jeans length is not a fundamental, universal constant, the same always and anywhere, but rather a physical value depending on concrete conditions, i.e. the pressure and density of the medium. It increases as the pressure increases or as the density decreases: recall the concept of the gravitational instability in the expanding universe presented in Chapter 2. If the physical conditions in a prestellar medium are known, we can estimate the critical Jeans length and thus find the size of condensations into which this medium could disintegrate: to be more accurate, we can find the lower bound of the size.

Physical conditions were essentially different during different epochs of star formation. The first stars in the Galaxy appeared in a contracting protogalactic cloud, and this star formation finally put an end to the contraction of the protogalaxy as a whole. The possibility of fragmentation of the protogalactic cloud required that its size greatly exceed the Jeans length, then the separate fragments of it could also have sizes exceeding the critical length. A protogalactic cloud in this sense is the infinite medium in Newton's concept: infinity always means, in fact, a considerable predominance of something over something. It is only in cosmology that the infinity of the universe as a whole has an absolute rather than a relative sense; if a volume of space is infinite, it is really infinite, i.e. cannot be expressed in terms of any finite values, and it is not simply much greater than any other volume.

The stars appearing in the Galaxy now are born in dense and cold gas-dust clouds which are greatly different in their properties from the layer protogalaxy. But here again the size of initial clouds should be much greater than the Jeans length for these clouds to break up into fragments.

Cascade Fragmentation

The protogalactic cloud did not break up into condensations with masses equal to those of stars; it is more likely that first more massive condensations were formed which, in turn, decomposed into smaller fragments until the condensations formed star masses. This successive fragmentation is evidently possible if the Jeans length gradually decreases in the process of general contraction of both the whole cloud and each of the fragments into which it breaks up (Fig. 34).

The Jeans length does decrease during contraction. And it is essential that it decreases faster than the total size of the whole cloud while it contracts. This is why fragments progressively smaller in size and mass can appear.

This behaviour of the Jeans length is controlled by the specific nature of change in the pressure and density in the contracting cloud. The fragments into which a protogalactic cloud breaks up emit photons into the environment and therefore lose some of their energy. However,

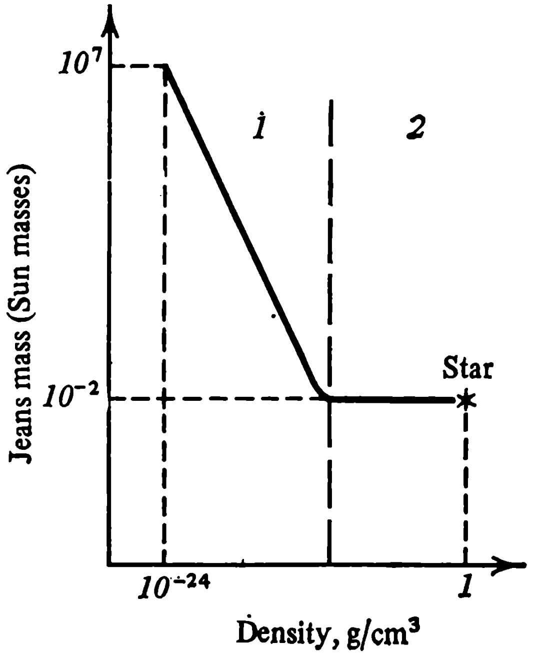

Fig. 34
The Jeans mass vs. density during cascade fragmentation. The Jeans mass is the amount of matter within a volume of the Jeans length size. Logarithmic scales along both axes. *1*—the area of transparency where contraction is accompanied by fragmentation; *2*—the area of opacity where contraction occurs without fragmentation

they do not cool because the thermal energy of each fragment is constantly replenished during contraction at the expense of gravitational energy. The result is that contraction occurs at an almost constant temperature close to ten thousand kelvins.

This temperature corresponds to the boundary between the ionized and nonionized states of hydrogen, the principal element of protogalactic gas. At greater temperatures, electrons and nuclei are separated from each other by chaotic thermal motion, while at lower temperatures, electrons and ions join each other, thus producing atoms which participate as a whole in the thermal motion. An ionized gas emits photons rather intensely because of the electric interaction between the moving electrons and ions. Furthermore, when the temperature is close to ten thousand kelvins, i.e. when neutral atoms appear in the medium, the thermal energy of particles is great enough to excite atoms in collisions and transfer an appreciable part of the energy of thermal motions to the electrons bonded to the atoms. This energy is not retained in the atoms for long: the atoms soon release the excess energy of their electrons as quanta of radiation, i.e. photons, which leave the condensation at a high rate. At lower temperatures, an essential emission cannot be produced by free electrons because most of them are bonded to atoms; at the same time the radiation of atoms cannot be intense because the energy of thermal motions becomes too small and collisions between particles cannot carry atoms into an excited state. And if the medium does not produce hydrogen molecules (these molecules require less energy to be excited by thermal motion), there is generally no emission of energy from a system at temperatures less than 10^4 K.

One may say that a self-regulation of temperature occurs in a compressing gas. The gas cools rather intensely to the marginal temperature of ten thousand kelvins, but further cooling is impossible. And any heating above this temperature because of contraction is immediately eliminated by radiation. Consequently, the temperature during the whole process of cloud contraction is maintained around the above-mentioned specific value.

Calculations show that the Jeans length is about 10^{23}-3×10^{21} cm at a temperature of 10^4 K and density from 10^{-27} to 10^{-24} g/cm^3 (the latter value corresponds to the typical density of a galaxy, i.e. to the density of a protogalactic cloud when its general contraction comes to a halt). A volume of this size contains masses from 30 to 1000 million Sun masses. These condensations probably appeared and were then contracted and fragmented at an almost constant temperature.

While a large cloud contracts and its density increases, the Jeans length decreases with time from the characteristic value 10^{23}-3×10^{21} cm, and the masses of fragments within this cloud also decrease. Is the process unrestricted, so that infinitely small masses can appear? No, it is not: sooner or later the fragmentation ceases and further separation of small fragments becomes impossible.

The point is that while the density of matter increases, it becomes more difficult for the photons which carry some of the energy from a condensation to leave this condensation. While they leave each of the fragments with initial, comparatively small densities without hindrance, the picture is different with greater densities at the later stages of contraction. The photons now, as it were, get stuck in the medium and move to the boundary of the fragment much more slowly than in a free flight because they interact with the atoms and electrons deflecting them from a rectilinear path. A photon leaves a transparent fragment along a straight line, whereas it leaves an opaque fragment along a broken line and does not get to the free surface quickly. Therefore, the whole transparent fragment loses all its energy at once, while the losses of energy of the opaque fragment occur only through its surface as the photons generated inside reach the surface.

As soon as a fragment becomes opaque, the relationship between its size and the Jeans critical length (the latter being controlled by pressure and density) changes. As we have already mentioned, the decrease in the Jeans length in transparent fragments occurs faster than the reduction of size due to contraction; but now, because the energy release decreases, the Jeans length declines much more slowly, and the size of the fragment can catch up with it. It is clear that further fragmentation is impossible once the Jeans length is not smaller than the size of the condensation. While the energy is still radiated, the Jeans length continues to decrease, but the size of the condensation "tunes up" to it. Therefore the contraction of the fragment continues but proceeds comparatively more slowly and is not accompanied by further fragmentation.

The "last" fragment in cascade fragmentation, whose size is comparable to the Jeans length, is a finite volume incapable of further fragmentation; we have mentioned it above in connection with Newton's concept of gravitational condensation.

The idea of cascade fragmentation terminated by opacity was suggested by F. Hoyle in 1953. His junior colleague M. Rees has recently shown that the "last" fragment possesses a mass equalling about one hundredth of the Sun mass. It is noteworthy that this value does not practically depend on the concrete mechanisms of photon emission or the processes responsible for the opacity of the medium. It is only assumed that the contraction at the stage of transparency occurs at a constant temperature of about ten thousand kelvins (as we have seen, this is reasonable), and then the mass of the "last" fragment can only be expressed through fundamental physical constants such as the gravitational constant, quantum Planck's constant, the velocity of light, and the mass of a hydrogen atom. The mass of the "last" fragment is close to the smallest star masses of astronomical observations.

The picture of cascade fragmentation causing the formation of the first stars in our Galaxy and in other galaxies is both elegant and simple, but it does not, naturally, encompass the entire diversity of the physical processes in a contracting protogalactic cloud. Therefore one should not think that every star of the first generation should possess a very small mass. Such factors as the turbulization of the protogalactic medium, collisions between clouds, shock waves in the medium, etc. should also be taken into consideration, and they make the picture much more complicated and involved. And although much remains to be studied in more detail, it is nonetheless clear that the first generation could include both small stars and stars with masses dozens of times greater than the mass of the Sun.

Both small and large stars begin as contracting condensations that do not undergo further fragmentation. They are protostars whose temperature gradually increases because when they become opaque, the transport of energy outwards is slow. Essential changes take place in their inner structure: their density does not remain uniform in bulk, and it builds up in the centre faster than at the periphery. The central area becomes denser and denser, and therefore hotter, and finally nuclear reactions begin. The nuclear energy release increases the internal pressure so much that it even becomes capable of balancing the gravitation. The general contraction of the protostar ceases, and the transfer of energy from the surface outwards is compensated for by the nuclear reactions in the centre. Thus a protostar condensation turns into a star. This last stage of gravitational condensation takes, depending on the star mass, from several million years for massive stars to several hundred million years for stars less massive than the Sun.

The Interstellar Medium

Most of the stars of the first generation might not have been very massive; but those scarce ones whose mass was 10 to 50 times that of the Sun could essentially have evolved and transformed in ten or a hundred million years, i.e. in less time than the duration of the general contraction of a protogalaxy, which takes 10^8 to 10^{10} years. Massive stars use up their nuclear energy sources at a far greater rate than the Sun. When hydrogen in a star has burned up and transformed into helium, heavier elements, up to iron, are synthesized and then the star may explode and throw the heavy elements, both produced earlier and in the process of the explosion, into the environment.

It is probable that during the first several hundred million years of the existence of our Galaxy or any other galaxy, the accumulation of elements heavier than hydrogen and helium could have taken place in general. The cosmic medium gradually became more and more abundant with such elements as carbon, oxygen, nitrogen, etc. Recall that the primordial matter in the universe, out of which the first stars were formed, contained 70-75 mass per cent hydrogen, 25-30 per cent helium, and an insignificant admixture of deuterium, lithium, and other light elements (10^{-4} per cent). However, the initial material for the formation of the stars of the next generation was a medium essentially rich in carbon and other heavy elements (1-2 per cent) thrown out by the exploded stars of the first generation. The role of this admixture for the process of star formation and evolution, for the appearance of the Sun, its planets, and the origination of life on the Earth is very important.

The gas left after the formation of the first stars gravitated to the centre of the system. If a protogalaxy possessed a great angular momentum, the precipitation of the gas at the centre gave way to its accumulation in the central plane of the galaxy when the force of gravitation became comparable to the effects of inertia. The flat subsystem of our Galaxy appears to have been produced this way.

Further star formation could only have occurred in the central areas of the Galaxy and in its plane: there were considerable masses of gas there rich in heavy elements. This process continues during the modern epoch in the spiral arms of the Galaxy, where observations reveal most of the youngest and brightest stars. It is noteworthy that the formation of a new generation of stars occurs, as it were, before our eyes, and therefore, in contrast to the problems discussed above, direct observational study of the cosmogonical process is possible in this case.

However, traditional optical observations, i.e. observations in the visible light by means of conventional optical telescopes, are strongly hindered by the fact that the Earth is in the disk of the Galaxy, and we view it from within. The presence of clouds of gas and cosmic dust in the disk and therefore around us interferes with our observations directed in the plane of the Galaxy: there are too many clouds in the line of vision, and they hide from us everything that takes place at great distances. Observations of the disk of the Galaxy in the optical range are only possible in the close vicinity of radius of about 2 kpc, i.e. 6×10^{21} cm. However, the diameter of the disk is 30 kpc, or 10^{23} cm, which is about 20 times greater.

The thickness of the gas and dust layer is 300 pc, i.e. about 10^{21} cm. This is why only optical observations af right angles to the disk's plane or at any rate at great angles to it (rather than along the plane) are effective; fortunately, the radiation from these directions is absorbed insignificantly.

Radio astronomy possesses much more effective tools to study the disk of the Galaxy. Radio waves in the centimetre and millimetre wavelength bands propagate freely in the disk and are virtually not absorbed in the clouds of gas and dust. Of special importance are radio astronomical observations at the 21-cm wavelength. During the 1940's, van de Hulst and I. S. Shklovsky drew attention to the fact that the neutral atoms of hydrogen in the diffuse medium of the disk of the Galaxy should emit 21-cm radio waves. This emission is associated with the transition of the electron in a hydrogen atom from one energy level to another around the ground, i.e. nonexcited, state of the atom. These levels are very close to each other and only differ in the reciprocal orientation of the spins, i.e. the intrinsic angular momenta of the nucleus and the electron: the spins are parallel at higher energy levels and antiparallel at lower energy levels. The intensity of emission at this wavelength indicates the concentration and temperature of neutral hydrogen. Radio emission of this nature was first registered in 1951.

Further investigations showed that neutral hydrogen is distributed up to 20 kpc from the centre of the Galaxy. The temperature of interstellar atomic neutral hydrogen amounts to about 100 kelvins. The average concentration in the disk is close to one hydrogen atom per cubic centimetre, which corresponds to 10^{-24} g/cm^3: a value comparable to the average density of the Galaxy as a whole. The width of the layer of neutral hydrogen is about 200-300 pc. A considerable portion of neutral hydrogen is concentrated in the spiral arms of the Galaxy. The distribution of neutral hydrogen emitted at 21-cm wavelength is crucial when the spiral structure of the Galaxy is discussed.

It is significant that the diffuse medium of the Galaxy contains molecules that also emit in the radio-frequency range. The hydroxyl group (OH) emits at 18-cm wavelength; the emission of interstellar hydroxyl was predicted in 1949 by I. S. Shklovsky and discovered in observations in 1963. The intensity of this emission in the direction from the centre of the Galaxy turned out to be rather considerable. Later about 50 different molecules were discovered as well, such as water (H_2O), hydrogen (H_2), carbon monoxide (CO), etc. Rather unexpected was the discovery of such polyatomic molecules as, for instance, ethanol (C_2H_5OH). The list of molecules discovered in the interstellar medium is supplemented with every passing year.

The distribution of molecules, or, more accurately, of clouds abundant in molecules, does not follow the distribution of atomic neutral hydrogen in the disk of the Galaxy. The molecules of carbon monoxide are mainly found within a ring around the centre of the Galaxy whose outer boundary is close to the orbit of the Sun (its radius is about 10 kpc), and the inner boundary has a radius of about 3 kpc. The greatest density of the molecules is observed between the radii of about 4 and 7 kpc, where the thickness of the ring amounts to 100 pc, which is two or three times less than the thickness of the neutral hydrogen layer.

The temperature of the medium in the areas where carbon monoxide and other molecules are observed is about 10 kelvins. Collisions can only excite the lowest energy states at such temperatures. These states in a carbon monoxide molecule are associated with its rotation about its own axis. The collisions at the above-mentioned temperature cause this rotation (the weakest one allowed for this molecule by the laws of quantum mechanics), and a stop, or cessation of this rotation, is accompanied by the emission of a quantum of electromagnetic waves, corresponding to the 2.6-mm wavelength. The emission of interstellar carbon monoxide was discovered in 1970 by A. Penzias and R. Wilson (five years earlier they had discovered relict radiation) together with their colleagues.

The most abundant molecules in the disk of the Galaxy are those of hydrogen (H_2). Molecular hydrogen was discovered in the interstellar medium by its ultraviolet lines of emission and absorption. This became possible only in the 1970's, when extraterrestrial astronomy was developed. (The ultraviolet radiation of all celestial bodies, except the Sun, is completely absorbed by the Earth's atmosphere.) The total mass of molecular hydrogen is probably close to that of atomic hydrogen in the disk of the Galaxy. This estimate is obtained on the basis of the data on the relative contents of molecular hydrogen and carbon monoxide in close molecular clouds where the total numbers of both molecules can be determined independently (the data of radio, X-ray, and ultraviolet observations from spacecraft are used for this). The number of hydrogen molecules in such clouds proves to be about ten thousand times greater than that of carbon monoxide molecules. Direct observations of molecular hydrogen are difficult at great distances, but scientists believe that the above relationship is typical for the molecular clouds of the galactic disk. If this is so, then both molecular hydrogen and carbon monoxide are concentrated mainly in the ring of molecular clouds; however, the general distribution of molecular hydrogen is different from that of atomic hydrogen. This is illustrated in Fig. 35.

The origin of molecules, especially the polyatomic ones, appears to be associated with the physicochemical processes in which the active role is played by cosmic dust. Cosmic dust consists of tiny hard particles, mainly carbon ones or with an admixture of ice, measuring 10^{-4}-10^{-5} cm. They can absorb visible light intensely, and this is, as we have already mentioned, the main hindrance to optical observations in the disk of the Galaxy. Dust accounts for about one per cent of all mass in the interstellar medium. Atoms collide with dust particles in their random motions, stick to their surface, and can then be involved in chemical interaction and join into molecules. In fact, the conditions on the surface of dust particles are favourable: the concentration of atoms is much greater here than in the interstellar medium on the average, and furthermore, the nonexcited state of the atoms provides enough time for the completion of chemical processes. The surface of the dust particles acts as both a flask and a catalyst. The interstellar medium thus produces very complex organic molecules, e.g. amino acids. Most likely, the origin of the organic compounds found in chondrites, i.e. meteorites of a specific type, is the same.

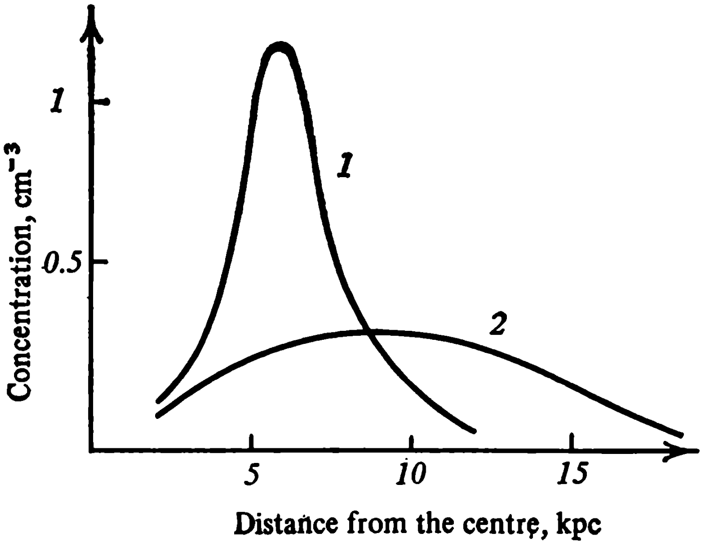

Fig. 35
The distribution of molecular (*1*) and atomic (*2*) hydrogen in the disk of the Galaxy.

The dust particles themselves appear to form in the extended atmospheres of red giants, cold stars abundant in heavy elements, mostly carbon. Matter escapes the atmospheres of these stars, diffuses in space, and thus cosmic dust particles mix with the gaseous matter of the interstellar medium.

The space density of the gas-dust medium varies more or less regularly only over large scales comparable to the size of the entire galactic disk. This large-scale distribution of hydrogen is shown in Fig. 35; cosmic dust, as stated above, associates with molecules and therefore follows their distribution. The distribution of gas and dust in smaller scales is very patchy and irregular; the areas of increased concentration alternate with almost void volumes or volumes filled with hot (up to a million kelvins) and very rarefied gas. The condensations, i.e. clouds, possess a density some ten times greater than the average density of the medium. A typical cloud 40 to 50 pc in size contains a mass of a hundred thousand or even a million Sun masses. The total number of such clouds in the Galaxy is 5-10 thousand.

The clouds move randomly with rather great velocities reaching 6-8 km/s. Furthermore, they naturally participate in the general rotation of the disk of the Galaxy; recall that the linear velocity of this rotation is 220-250 km/s near the Sun.

The present-day formation of stars occurs in these condensations, comparatively cold and dense clouds. Before discussing this point, let us consider some other properties of the interstellar medium also essential for the mechanism of star formation.

The interstellar medium in the disk of the Galaxy possesses magnetic fields. Although their existence had been guessed long ago, the interstellar magnetic fields were directly found and investigated during the past two or three decades. Magnetic fields are detected by their influence on the properties of radio waves emitted by atoms and molecules within these fields. The fields also affect the propagation of radio waves in the interstellar medium (the plane of polarization of an electromagnetic wave rotates when it travels through a magnetized medium).

Magnetic fields are strongly associated with the clouds of the interstellar medium and move together with them. The general orientation of magnetic lines of force coincides with that of the arms in the spiral structure of the Galaxy. The magnetic field strength in the disk of the Galaxy reaches 3×10^{-6} to 3×10^{-5} oersted (Oe), which is several hundred thousand times less than that of the Earth.

The first ideas concerning the existence of interstellar magnetic fields were offered in connection with the problem of the cosmic rays retained in the Galaxy. Cosmic rays are charged particles (electrons, protons, the nuclei of heavier elements, etc.) moving with velocities close to the velocity of light. (See the book by V. L. Ginzburg (1967) listed in Recommended Literature.) Researchers believe that they travel within the Galaxy for several million years, although they could escape it in a hundred thousand years if allowed free flight. They are retained in the volume of the Galaxy by magnetic fields, which make charged particles move along magnetic lines of force following spiral or helical trajectories. The structure of the magnetic lines of force is complex shaped, so the cosmic rays have to loop in the volume of the Galaxy before they leave it (Fig. 36). This is strongly evidenced by recent observations of interstellar magnetic fields.

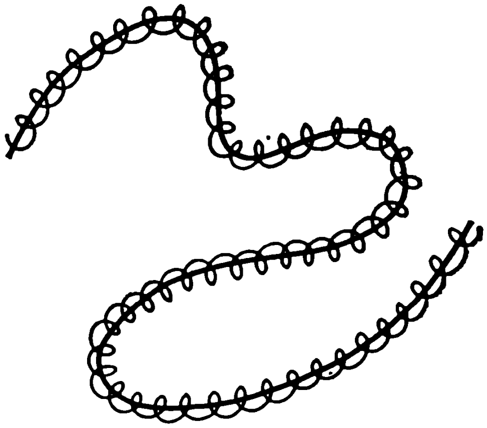

Fig. 36
The motion of a charged particle along a line of magnetic force.

Travelling in the interstellar magnetic fields, the fast electrons of cosmic rays emit radio waves that are, in point of fact, observed as the general radio emission of the Galaxy. The picture of the Galaxy in radio waves of the metre range is composed of three principal structural elements: the spherical (or maybe slightly flattened) halo, the disk, and the galactic nucleus. The sizes of the radio halo and the radio disk are close to those of the stellar halo and the stellar disk, respectively; the area of more intense radio emission in the centre (the nucleus) has a flattened shape and seems to follow the shape of the galactic disk with the size (in the disk's plane) of about 300 pc. The nature of radio emission from the halo and the nucleus is the same as that of the emission from the disk: this is the emanation of cosmic ray electrons in a magnetic field. It follows that a magnetic field (possibly weaker than that in the disk) can be found in the galactic halo, while the magnetic field in the galactic nucleus is significantly stronger.

There are three components of the interstellar medium: gas-dust clouds, cosmic rays, and the magnetic field, and there is a specific kind of equilibrium between the three of them in which the total energy of the cosmic rays, the energy of the magnetic field, and the kinetic energy of the random motion of the clouds are approximately equal to each other. So far it has not become quite clear how this energy balance is established and maintained. However, there is no doubt there is an intense interplay between the clouds, the field, and the cosmic rays; thus, the cosmic rays are rather strongly "glued" to the lines of magnetic force, which, in turn, are strongly "tied" to the moving clouds. For these relationships to be really effective and, above all, reciprocal, the energies of the three components of the medium should be comparable. Each of the energies in question amounts to 10^{-19} J per cubic centimetre of the volume of the Galaxy on the average.

Recently, researchers have revealed that similar structure can be found in flat subsystems of other galaxies that are rich in diffuse medium and young bright stars. For instance, the nearest gigantic spiral galaxy of the Andromeda Nebula also contains a more or less uniform layer of neutral hydrogen and a ring of gas-dust clouds containing interstellar molecules. The brightest stars forming the spiral pattern of this galaxy are within this ring. A radio halo is also observed in the Andromeda galaxy.

There is a lot of gas and dust in irregular galaxies, such as the Large Magellanic Cloud observed in the sky of the southern hemisphere. There is also a great number of very young bright stars, which is an explicit indication of the relationship between star formation and the required presence of rarefied diffuse material.

The interstellar gas is almost absent only in elliptical galaxies. But there are no young stars there, and the process of star formation seems to have come to an end long ago. These galaxies possess only a spherical subsystem (typically flattened to a certain extent), and this subsystem is similar to the halo of our Galaxy in its structure and star population. Evolutionary processes in elliptical galaxies occur very slowly, the process is almost suspended, their stars take a thousand million years to change; and "life" goes on only in the most central parts of these galaxies. However, these areas show such activity which is not found in both spiral and irregular galaxies (see Chapter 5).

Young Stars

At the beginning of this chapter we pictured the formation of the first stars in the Galaxy. This picture is almost entirely based on theoretical considerations. However, there are direct and comprehensive observational data on the new generation of stars forming in the present-day epoch.

Observations definitely indicate that the youngest stars are located where there are large masses of rarefied diffuse matter. Newly born stars are found in dense and extended molecular clouds. There are also protostars submerged into condensations of gas and dust which produced them. Thus, astronomers believe that a rather uncommon T Taurus star and some other similar stars are actually protostars, i.e. dense condensations heated during contraction owing to the potential energy of gravitation rather than to nuclear reactions. The luminosity of protostars is almost the same as that of stars of the same mass; however, they are noticeably larger and cooler, especially the surface layers, and therefore the light of protostars is redder than that of the already formed stars. Stars of the T Taurus type are usually submerged into dark nebulae.

The effect of "cosmic masers" is a striking feature of protostars. Physicists have spent much effort inventing and producing these devices, and they were found to operate naturally in the cosmic medium. Recall that the maser effect is due to the so-called state of activation of the medium through which a source emits, i.e. the state of anomalously great number of excited atoms or molecules. The activation, or energy pumping, of the medium is carried out by some other source. Photons with an energy corresponding to that of a transition of an atom or a molecule from an excited level to the ground level are bound, while propagating in this medium, to stimulate the emission of new photons of the same energy, forcing atoms and molecules to return to the nonexcited state. The resulting flow of photons of the given energy can therefore prove to be strongly amplified. (Recall that the word "maser" is derived from "microwave amplification by stimulated emission of radiation".)

Such an unexpectedly intense radiation from the molecular clouds in the Great Orion Nebula was discovered in 1965 during radio astronomical observations at the 18-cm wavelength corresponding to the transition of a hydroxyl molecule from the excited state to the ground state. I. S. Shklovsky suggested that this radiation is due to a maser mechanism operating either in the dense clouds evolving into protostars or in the surface layers of protostars themselves. Observations revealed that the radiation comes from rather dense areas of the medium where the concentration of particles, apparently molecules for the most part, amounts to 10^8-10^9 per cubic centimetre. The size of the radiating areas is about a hundredth of a parsec. The temperature of the radiating matter is estimated at about 1000 kelvins. It appears that these should be the conditions in a protostar envelope.

The discovery of cosmic hydroxyl masers was followed by the discovery of water vapour masers emitting radio waves at 1.35-cm wavelength. The intensity of "water" masers proved to be greater than that of hydroxyl masers.

The source of energy pumping needed for the amplification of emission has not yet been revealed (probably, the energy is provided by the sufficiently intense infrared radiation of heated dust or of the protostar itself); however, the hypothesis of the protostar envelopes as the medium for the maser effect is supported by the fact that the vast molecular clouds where maser sources are observed always contain newly formed bright stars, a reliable indication of the continuing process of star formation.

The transformation of a protostar into a proper star is accompanied by significant changes in its environment. This is primarily associated with the influence of star radiation, which shifts into shorter wavelength range. As the surface of a star is heated, the colour of its radiation changes from red to blue, and photons belonging to the ultraviolet part of the spectrum appear. (An iron rod, when it is heated, also becomes first red, and then, with a further rise in the temperature, turns blue.) The "hot" photons break up molecules, and the ultraviolet radiation dissociates hydrogen molecules into atoms. Later, photons of still shorter wavelengths appear, and hydrogen atoms dissociate into protons and electrons, so the gas becomes ionized.

Ionized hydrogen envelops the star and forms a layer called the H II region. Photons heat the ionized gas while giving their energy to electrons and protons, so the gas temperature in the H II region reaches ten thousand kelvins. Since the gas here is much more heated than in the cloud surrounding this region (where the temperature is no more than 10-100 kelvins), the pressure in the region of ionized hydrogen is greater than beyond it. This creates a force which makes the hot region expand until the pressure within it is equal to the external pressure,

The H II regions emit visible light and therefore have been long since known to astronomers. Their relation to young stars has been soundly established. Most often, each region of this kind contains several hot young stars rather than only one. These stars are massive and bright. They belong to classes O and B in the Henry Draper system containing 9 classes designated by the letters O, B, A, F, G, K, M, R, and N in the order of changing the colour from blue to red. (The Sun belongs to the intermediary class G; the eye is most sensitive to the middle greenish-yellow portion of the solar spectrum.) According to every indication used to judge the age of a star, the stars belonging to classes O and B are the youngest in our Galaxy. (The reader can find more detail on the stellar colours and spectra and the relationships between the mass, age, and luminosity of stars in the book by B. A. Vorontsov-Vel'yaminov (1985) listed in Recommended Literature.)

A striking feature of classes O and B stars is that most of them are joined within groups of up to several hundred stars. These groups are called OB associations. V. A. Ambartsumyan was the first to suggest that this is an indication of the fact that stars are formed in groups, collectively, rather than individually. Moreover, according to recent data, young stars practically always join into groups of a kind. The largest groups of them are the gigantic star complexes discovered by Yu. N. Efremov. Almost every OB association known in our Galaxy belongs to star complexes or very young star clusters (the latter are typical members of star complexes). Most commonly, the age of the stars in star complexes does not exceed 50 million years.

There is no doubt that star complexes are formed in large molecular clouds, i.e. major condensations of gas and dust. There are molecular masers (protostars), OB associations, and H II regions observed within them. It is sometimes possible to follow the sequence of events developing in a molecular cloud and relating protostars, OB associations, and H II regions in a single evolutionary chain.

The most remarkable discovery in this field has recently been made by A. Blaaw (the Netherlands). He established that OB associations consist of subgroups of 5 to 20 stars each, and these subgroups are arranged within the volume of an association according to their age: a subgroup of the youngest members in the association is located at one end of the association, while the oldest members comprise a subgroup at the opposite end. This arrangement of subgroups by age implies that the process of star formation that brought forth the association occurred in successive "bursts", and a burst in one area triggered a burst in the adjacent area. A wave of star formation, as it were, propagated along the cloud and first produced stars at one end and then gradually proceeded to produce stars at the other end.

B. Elmegreen and C. Lada found a simple and convincing cosmogonical mechanism of the phenomenon discovered by A. Blaaw. As a starting point they assumed that the area of hot and ionized hydrogen, the H II region, developing around a newly born star of class O or B, expands and pushes out the surrounding cold gas. The velocities thus communicated to the cold gas amount to 5 or even 10 km/s, which is greater than the velocity of sound in this gas (no more than 1 km/s). Therefore a shock wave appears in the cold molecular gas, which compresses and heats the gas beyond it. Gradual radiation cooling of the gas causes its further compression, and several million years later the temperature and pressure are bound to fall (and the density is bound to increase) so that the gravitational condensation of the layer would become possible.

As mentioned at the beginning of this chapter, condensation is accompanied by a cascade fragmentation of matter. This does not occur in quite the same way in the gas-dust medium of molecular clouds as it occurs in the primordial gas of a protogalaxy. For instance, the temperature cannot be maintained at the level of ten thousand kelvins; since there is an admixture of carbon, oxygen, and nitrogen, the temperature decreases by a factor of several hundred times. But probably this is why this process can produce mostly stars with greater masses. Young massive stars always belong to classes O or B. And this means that they produce an H II region around them in turn; the new H II region creates a new shock wave in fresh molecular gas, and this shock wave is bound to produce a new burst of star formation in several million years. The process repeats itself, and the focus of star formation moves from one end of the cloud to the other.

This picture gives a demonstrative explanation for the very existence of star subgroups of different ages within the same association and for the age sequence in the arrangement of the subgroups.

Moreover, direct observational data have recently appeared concerning waves of star formation propagating in large molecular clouds. For instance, a bright nebula in the Cassiopeia constellation is a luminous cloud of ionized hydrogen with young stars submerged into it; as radio observations showed, the nebula is a part of a major molecular cloud 50 pc in size. The Cassiopeia Nebula itself (Fig. 37) includes two H II regions. The extreme eastern region (it is designated as IC 1805) is a rather old diffuse H II region which is expanding and becoming diffuse, exposing within itself a subgroup of at least 20 stars of classes O and B; to the west of it there is a younger and less rarefied H II region (it is designated as IC 1795) which is also expanding. No doubt, the latter contains young stars, but they are screened from us by clouds of gas and dust. The western boundary of the young H II region moves into the molecular cloud. Near this boundary there are several very compact sources of radio waves and infrared radiation: these are either massive protostars or newly formed massive stars. This is precisely what a wave of star formation in a molecular cloud should look like.

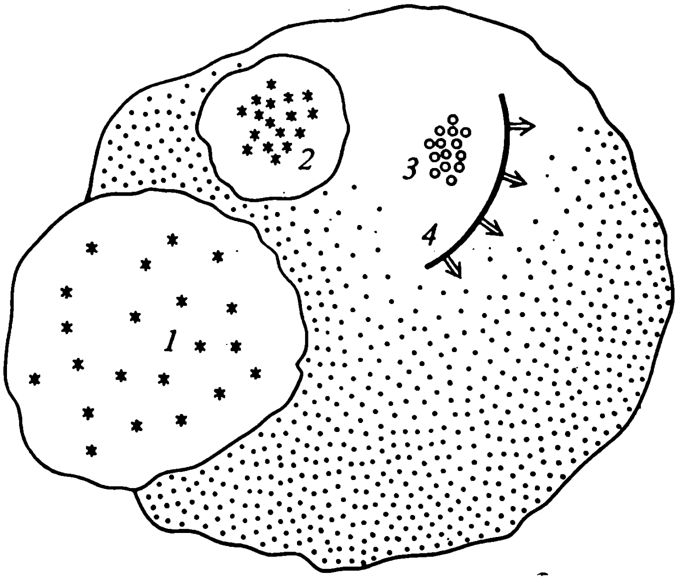

Fig. 37
A wave of star formation propagating from east to west in a molecular cloud. *1*—an "old" H II region with an "old" subgroup of classes O and B stars; *2*—a "young" H II region with a "young" subgroup of classes O and B stars; *3*—protostars; *4*—a shock wave front. Astronomical maps show the east to the left and the west to the right.

But what produced the first bright stars which began the "chain reaction" of star formation? No doubt, this is associated with an external influence on the molecular cloud. Possibly, it collided with another molecular cloud because the motion of clouds is chaotic in the disk of our Galaxy. The velocity of colliding clouds is commonly greater than the velocity of sound in their matter; therefore such a collision could result in shock waves at the contacting sides of the clouds. More accurately, there should be two shock waves propagating in opposite directions from the surface of contact; behind each of the shock wave fronts there appears a layer of compressed gas capable of further gravitational condensation and fragmentation. However, calculations show that the probability of such a collision is rather small: the rate of expected collisions in our Galaxy is hardly more than one per ten million years. This appears to be insufficient to explain the observed rate of star formation in molecular clouds.

Another possibility is related to star explosions during the later stages of their evolution. Such explosions have been observed both in our Galaxy and in other galaxies as bursts of supernovae. The envelope of such a star or sometimes the whole of it blows up with a great velocity and creates an explosion shock wave. If this occurs near a neighbouring molecular cloud (within several parsecs), a condensation may occur in the cloud sufficient enough to "ignite" the process of star formation in it.

Finally, shock waves can form in molecular clouds owing to the spiral wave of density in the disk of the Galaxy. The rotation of the disk of the Galaxy turns the perturbations propagating in it into spiral waves (see Chapter 3). These waves "surge" over the cool gas of the clouds with a velocity exceeding the velocity of sound in the gas and therefore generate shock waves in the clouds. Astronomers noticed that H II regions are found more commonly in the molecular clouds occurring along spiral arms rather than in those outside the arms. Possibly, this is precisely because shock waves are excited in the clouds by spiral waves and produce hot stars in turn.

It is also possible that the three mechanisms under consideration operate under the actual conditions of the interstellar medium; at any rate, the available data do not allow scientists to give a definite preference to one of them.

Another problem in connection with the picture of the chain reaction of star formation is, perhaps, more difficult. Why are not only massive stars of classes O and B but also a greater number of common stars like our Sun formed? These stars are also produced by the gravitational condensation and fragmentation of matter, and they probably appear in the same "bursts" of star formation that give rise to massive stars. The fragmentation of a major condensation into large fragments is hardly possible without smaller fragments, and there may be many more of them than of massive fragments.

The process of fragmentation producing a large enough number of fragments should obey very general statistical laws independent of both the nature of fragmenting bodies and the concrete mechanisms of fragmentation. During the 1940's, A. N. Kolmogorov drew attention to the fact that, according to numerous empirical data, the sizes and masses of gold particles found while washing sand from gold dust are always distributed according to a definite (lognormal) law. The majority of particles possess an average mass. The greater the deviation of the mass of particles from their average mass, the smaller the number of both less and more massive particles. A. N. Kolmogorov proved that this distribution applies to numerous and most diverse processes of fragmentation, or cascade fragmentation as in our case, when the initial mass is successively fragmented into ever smaller parts. Both the initial mass and any fragment may separate in any act of fragmentation into a random number of fragments arbitrarily distributed by masses. It is only necessary, and this is a rather strict requirement, that the probability of fragmentation in each process should not depend on the initial mass.

Astronomical data on star masses and both theoretical and observational data on the rate of their evolution allowed E. Salpeter (USA) to suggest an empirical stellar mass distribution law. According to this law, the greater a given mass, the fewer stars within it (Fig. 38). In contrast to Kolmogorov's distribution, there is no typical, average mass. However, it probably does not apply to stars of the smallest masses: their number appears to be less than that expected according to Salpeter's law, and the actual distribution should "tend" towards smaller masses. Despite this tendency, the real distribution of star masses is hardly thought to be strictly homogeneous with respect to a typical mass. A comparison of these facts and Kolmogorov's theory suggests that there seems to be a definite dependence of the fragmentation probability on the mass of fragments (i.e. the requirement assumed in the mathematical theory is not satisfied during star formation). This dependence should be valid if the sizes of fragments in each process of fragmentation slightly exceed the Jeans length.

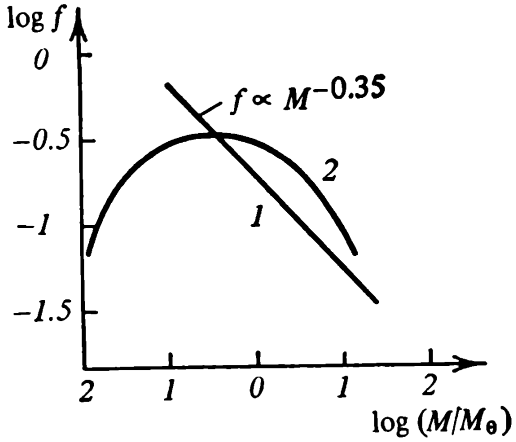

Fig. 38
The mass distribution of stars: f is the fraction of the total mass of stars per unit logarithmic interval of star masses (measured in Sun masses $M_{\odot}$). *1*—Salpeter's law; *2*—Kolmogorov's law.

The scheme of cascade fragmentation we have discussed while considering the first stars in our Galaxy is also applicable to the new generation of stars, although concrete physical processes (heat transfer and cooling and heating of gas) are different, in this case. A consistent theory explaining both typical masses of stars and their mass distribution should combine the ideas of cascade fragmentation with the comprehensive results of the probability theory, such as Kolmogorov's law. So far such a theory has not yet been elaborated. But it is noteworthy that the cascade fragmentation in star formation occurs, in fact, so that the sizes of fragments are close to the critical Jeans length, and therefore it can be expected that the stellar mass distribution law should not be homogeneous with respect to a typical mass, i.e. it should rather follow Salpeter's law.

Instabilities and Clouds

Thus, the stars of new generations are formed in molecular hydrogen clouds rich in other molecules and cosmic dust. But what is the origin of these clouds?

The physical state of the interstellar medium is controlled primarily by the processes of heating and cooling. The heating is due to cosmic rays (more definitely, to their particles of comparatively low energy) and to the background X-ray radiation of the universe. This radiation appears to be emitted by such sources of X-rays as clusters of galaxies. The interstellar gas in the vicinity of hot bright stars may be heated by their radiation, for instance, as it happens in H II regions. The cooling is mainly due to collisions of atoms and molecules. When they collide, they are excited by the energy of their chaotic motion; when they return to the initial nonexcited state, they emit the acquired energy as photons which leave the medium. Consequently, the thermal energy of particles is transformed into radiation and eventually lost by the medium.

The energy received from the outside by each element of the medium per unit time while it is heated is proportional to the number of particles in a given element; but the loss of energy is proportional to the square of the number of particles. The difference is related to the fact that the heating is carried out by "alien" particles, e.g. cosmic rays or X-ray photons, while the cooling requires pair collisions between the gas particles. Therefore, only one gas particle participates in a single process of heating, while two of them participate in a process of cooling, and hence the rate of heating is proportional to the first power of the number of particles, while the rate of cooling is proportional to the second power of the number of particles.

These processes of heating and cooling make the medium unstable: it cannot remain uniform but tends to break up into condensations surrounded by rarefied gas. This kind of instability in its final manifestation is very much like gravitational instability which we have discussed in great detail, but its mechanism is not related to gravitation and is entirely controlled by thermal processes. Therefore this kind of instability is said to be thermal.

To understand the nature of thermal instability, let us suppose that the interstellar gas is in such a state that its density is uniform everywhere, while the heating of each element of the medium is perfectly balanced by its cooling. Now suppose that an element of the gas becomes slightly denser than its environment, i.e. the number of particles in it is somewhat greater than in an equal volume of the uniform medium. It is evident that the balance between heating and cooling becomes immediately shifted: now both processes occur faster than in the surrounding medium, but cooling is more intense than heating because cooling is more sensitive to the number of particles. Therefore the temperature of the gas in the element drops, and this leads to a drop in the pressure in it. So the external (greater) pressure of the medium compresses the element in order to restore the pressure in it. Now the density and hence the prevalence of cooling over heating increase, the temperature and pressure decrease, and the external pressure compresses the element still greater. Thus the compression, once started, intensifies. The characteristic feature of instability is evident in the phenomenon: any slight condensation in the interstellar medium spontaneously intensifies rather than smoothes out.

Thermal instability stops developing when the gas in the compressed element cools to such a degree that the thermal energy of its particles proves to be insufficient for the excitation of atoms and molecules. Then the cooling of the element ceases, and its equilibrium with the surrounding medium is established. Although the temperature in such a condensation is lower than that of the environment, the density is greater, and therefore the pressure (it is proportional to temperature times density) finally becomes equal to the pressure in the surrounding medium. The external pressure restores the initial pressure in the element, but now the temperature is lower and the density is greater.

Thus cold and dense clouds of the interstellar gas appear to be surrounded by the hotter and rarefied medium. According to theoretical estimates made by G. Field, S. B. Pikelner, and S. A. Kaplan, the clouds should have the masses and sizes as calculated by observational data. The pressure equals that of the surrounding medium, and this makes these clouds stable. They can only be destroyed by the star formation initiated by external effects.

Besides thermal instability, there are other processes operating in the interstellar medium, also causing nonuniformity and general patchiness in it. One of them is associated with magnetic fields present everywhere in the gas of the disk of the Galaxy. Their lines of force are generally parallel to the plane of the Galaxy; magnetic field strengths amount here to several microoersteds.

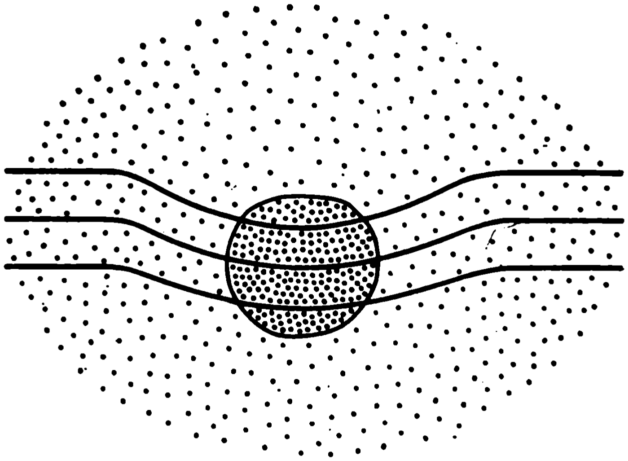

Fig. 39
The instability and formation of a cloud in a medium with a magnetic field.

A layer of gas with horizontal lines of force possesses additional elasticity: the magnetic field impedes the motion of the gas at right angles to the lines of force. This "magnetic elasticity" is only effective across the direction of the field, while the gas moves freely along the lines of force.

If the lines of magnetic force are bent in a volume of the medium as shown in Fig. 39, the gas "slides down" into the trough under the force of gravitation directed to the central plane of the Galaxy (downwards in Fig. 39). The kink becomes greater under the effect of these new portions of gas, the trough deepens, and the flow of gas into it increases. Consequently, a cloud, i.e. a considerable condensation, is created in the medium. This process has been shown by E. Parker to be very effective in the gas of the spiral arms of the Galaxy where the lines of magnetic force are mainly concentrated.

And last but not least, the interstellar medium may develop gravitational instability making the clouds still denser. As always, the size of the cloud in this case should be greater than the critical Jeans length. This prerequisite does not seem to be satisfied in minor clouds of the regions of nonionized atomic hydrogen revealing themselves in the 21-cm wavelength radiation. However, major clouds in the ring of molecular hydrogen in the disk of the Galaxy are comparable in size to the critical length corresponding to their density and temperature. This probably means that gravitational instability is either operating right now or operated earlier, shaping the clouds as they are currently observed.

The Life of a Star

The intensification of small deviations from uniformity in diffuse matter, the appearance of clouds, and their gravitational condensation and fragmentation are the aspects of the process eventually leading to the formation of stars and star complexes. Although the concrete physical processes in the primordial hydrogen-helium matter of a protogalaxy and in the present-day interstellar gas are different, the succession of events is identical both for the first stars of our Galaxy and for the stars of new generations. And stars themselves, both those formed thousands of millions of years ago and those appearing during the recent epoch, are very similar to each other in their principal features.

The stars of any generation, both in our Galaxy and in other galaxies, start their evolution from a state like the current state of our Sun. The differences in stars of different generations result from their further evolution, whose rate and the final outcome radically depend on the mass of a star. Certain features of the structure and evolution of a star are also associated with the chemical composition of a star: the contents of elements heavier than hydrogen and helium steadily increase from one generation to the next. However, the most important factor is the mass of the star. As we have already had a chance to see, the mass determines first of all the duration of the initial, Sun-like state of the star, when it emits energy owing to internal nuclear transformations.

A star, whose mass is several dozen times that of the Sun (e.g. a class O star), maintains this state for no more than 3-8 million years. A star with a mass equal to that of the Sun can prolong this period for another 13,000-15,000 million years. A star with one half of the Sun's mass remains in the initial state for about 100,000 million years. Compare these data with the age of our Galaxy: some 12,000-15,000 million years have passed since it was formed. It logically follows that only the least massive stars of the first generation, those with masses less than the Sun's, can be observed now in their initial state. These stars compose the oldest population of the Galaxy, that of its spherical subsystem. There were brighter and more massive stars there a long time ago, but they no longer exist.

Having exhausted the resources of hydrogen in the central region in several million years' time, a massive star enters a new phase of its evolution. Now nuclear reactions occur in a narrow layer of hydrogen around the core of the star rather than in the core, which consists entirely of helium produced by hydrogen nuclear "burning". A star with this kind of a "layer" source of energy increases its luminosity and "swells". Since the star becomes greater in size, the temperature of its surface layers declines, and therefore its colour turns from blue to red. Consequently, a class O or B star transforms into a star of another type, a red giant.

At the same time some changes take place in the helium nucleus of the star. It becomes more heated and gradually contracts. When the temperature reaches 100-150 million kelvins, helium starts burning and nuclear reactions occur, fusing three nuclei of helium into a carbon nucleus. The associated energy release raises the temperature of the core, and the increased pressure halts contraction. The total luminosity of the star becomes greater since now both the layer source and the helium nucleus contribute to it. Calculations show that this leads to a "small" rise in the surface temperature of a star. However, this stage is short, shorter than the initial "solar" stage, and ceases when there is no more helium in the hot and dense nucleus.

Now what happens to the star? It loses its envelope, i.e. its outer layers leave the core and expand, thus producing a planetary nebula. These interesting objects have long since been observed by astronomers. (The reader can find more detail on them in the book by I. S. Shklovsky (1982) listed in Recommended Literature.) As to the core, if its mass is no more than 1.2 Sun masses, it is exposed and becomes a small-size star with a rather high temperature. Such stars are said to be white dwarfs: white because of their colour that corresponds to the high temperature of the surface, and dwarfs because of their weak luminosity. Their luminosity is weak since the radiating surface of the star is small: the radius of a white dwarf is some hundred times less than that of the Sun, i.e. it is comparable to the radius of the Earth. Emitting light and gradually cooling, these stars become invisible in a thousand million years or so because their thermal energy is spent completely. There are no other sources of energy in them, and this is actually their steady state, that of a white dwarf turned into a black dwarf; it is the final stage of evolution of most stars.

Something like this is expected to happen to our Sun. It is going to turn into a red giant in 8000-9000 million years, then it will shed its envelope and become first a white dwarf and later a black dwarf.

The fate of a more massive star is different. Having exhausted its resources of nuclear fuel, it is also capable of shedding its envelope, but this occurs as a powerful explosion rather than a gradual process. Probably, these are the supernovae, the stars that suddenly burst into very great brilliance as a result of their blowing up. The luminosity of a supernova increases by several hundred million times in a matter of a few days, and then the star remains brighter than a whole galaxy for a week or a month. The envelope of the star rapidly expands and creates a shock wave in the interstellar medium (which can "ignite" a wave of star formation; see above). The rest, i.e. the core of the star, contracts quickly, and if its mass does not exceed two Sun masses, it becomes a neutron star.

The density of a neutron star is comparable to that of atomic nuclei, i.e. 10^{15} g/cm^3. This density is produced by the star's gravitation, because its mass does not allow it to remain in the state of a white dwarf. The radius of a neutron star is about ten kilometres, which is approximately a hundred thousand times less than the radius of the Sun, or at least six hundred times less than the radius of the Earth, i.e. the size of a small city.

Bursts of supernovae have been known since ancient times; a colourful description of such a phenomenon that occurred in our Galaxy in A. D. 1054 has been found in Chinese chronicles. The "guest star", as it is called in the chronicles, could be seen in the daytime, and its brilliance was only surpassed by the Sun and the Moon. The envelope of the supernova shed at the time is present in the sky today: this is the famous Crab Nebula studied by astronomers for over two centuries. A neutron star, the remnant of the supernova's burst, was discovered in the centre of the Crab Nebula in 1967.

Bursts of supernovae are rare phenomena in the present-day state of the Galaxy: they do not occur more frequently than once in a hundred years on the average. The latest supernova in our Galaxy was observed by J. Kepler in 1604. Bursts of supernovae in other galaxies take place about as often as in ours, so while observing several hundred galaxies for a year, we can certainly register such an event.

The neutron star in the Crab Nebula was found to be a source of short radio-wave pulses with a regular period of 0.033 s. This is an unusually small period; variable stars have long since been known, there are stars with regular changes of brightness, but only pulsars discovered in 1967 possess such small periods. The neutron star in the Crab Nebula is a pulsar, one of the first ever discovered. The first one was the pulsar with a period of 1.33 s discovered by a group of astronomers headed by A. Hewish. More than a hundred pulsars have been registered to date, and the periodicity of the pulse train is close to one second for the overwhelming majority of pulsars.

The periodicity of pulsar radiation is related to their fast rotation: such a star emits a narrow beam of radio waves like a beacon does. The beam periodically comes into the line of vision, and thus an observer sees a regular train of pulse radiation. However, none of the common stars, nor the Sun, nor a much denser white dwarf could rotate with a period characteristic for pulsars: any of them would immediately be torn apart by inertia. Only a neutron star can withstand and rotate as a top without being destroyed by inertia: such a star is very dense and compact enough for that. In fact, this kind of reasoning has been the primary and most convincing argument suggesting that pulsars are neutron stars.

Pulsars that emit mainly X-rays rather than radio waves were discovered during the 1970's. They proved to be neutron stars included in binary systems of stars. We shall treat some interesting phenomena occurring in close binary systems below, but now we are going to return to the fate of a massive star that has no more resources of nuclear fuel and examine the physical processes in white dwarfs and neutron stars.

The greater the mass of a star, the greater its gravitation. The gravitation tends to compress the star, and if the temperature and pressure in it are no longer maintained by nuclear energy release, then nothing hinders the action of gravitation. However, while a star becomes denser owing to gravitational contraction, its matter develops a special kind of elasticity unassociated with the temperature and the usual kind of pressure due to the thermal motion of particles. Even at absolute zero, matter possesses a nonzero elasticity whose nature is in no way related to thermodynamics. These are the conditions at which the effects of quantum mechanics operate, and they create the effective pressure of matter particles. (The reader can find more detail on the properties of superdense matter in the book by A. S. Kompaneets (1976) listed in Recommended Literature.) The gas is said to be degenerate in such a state, and when the density increases, the gas particles of smaller mass reach the degenerate state sooner.

Electrons are first to become degenerate in the mixture of electrons and nuclei of which a contracting star consists. Additional elasticity appears because of this, and it halts the contraction of the star if its mass does not exceed 1.2 Sun masses. This critical mass was calculated by S. Chandrasekhar. A star in which gravitation is balanced by the quantum-mechanical pressure effect of the degenerate electron gas is a white dwarf. It is clear that the cooling of a white dwarf cannot influence the pressure of degenerate electrons: it is going to be the same at absolute zero.

If the mass exceeds the Chandrasekhar limit (1.2 Sun masses), the pressure of degenerate electrons is not sufficient to counteract the gravitation and halt the contraction. The contraction of a star continues, the density becomes greater and greater, and the composition of the star matter changes: electrons, as it were, are pushed into nuclei and fuse with protons, turning them into neutrons. The fusion of electrons with protons is accompanied by the emission of a large number of neutrinos; they leave the star freely, withdrawing all the energy released during the contraction owing to the potential gravitational energy of the star.

But the contraction continues, and now neutrons become degenerate. The additional pressure in the neutron gas is capable of halting the contraction of the star and balancing its own gravitation if the mass of the star does not exceed two Sun masses. This is how a superdense star appears, which is said to be a neutron star because it consists mainly of neutrons and does not contract owing to their quantum-mechanical pressure effect. A theory of neutron stars was developed during the 1930's by L. D. Landau long before their discovery. They were thought then to be purely hypothetical objects, but their existence in nature inevitably followed the laws of physics.

Note that the fast rotation of neutron stars, making them pulsar beacons, is a direct consequence of their enormous contraction. Every star rotates, and our Sun rotates with a period of about a month (more accurately, the period of rotation at its equator is equal to 26 days). When a body contracts, its size decreases (recall a figure skater: as he or she moves the arms nearer to the body, the rotation becomes noticeably faster). The period of rotation declines in proportion to the square of the size. If the Sun were contracted to the size of a neutron star, it would rotate with a period of about one second, as is typical for pulsars.

It is essential that while a star contracts, not only the velocity of its rotation but also its magnetic field increase: the latter intensifies in inverse proportion to the square of the radius. When the densities typical for a neutron star are reached, the magnetic field may become

10,000-100,000 million times stronger than in the initial state of the star. The average strength of the Sun's magnetic field is about 1 Oe; if it were contracted to have a radius of a neutron star, the strength of the field would be equal to 10^{10} Oe. As far as can be deduced, the magnetic fields of pulsars are even stronger, up to 3×10^{12} Oe. It is precisely the magnetic fields which shape the beam of the radio waves emitted by a pulsar, and this beam rotates, as does the beam of a beacon, with the frequency of rotation of a neutron star. The joint action of rotation and the magnetic fields determines the exceptionally regular timing of the pulses with a period equal to the duration of the star's rotation cycle, and this period is maintained with an unbelievable accuracy: to the eighth decimal place. This refers to both radio and X-ray pulsars.

What is the future of a star whose mass exceeds two Sun masses? The gravitational forces in this case are so enormous that neither thermal nor quantum-mechanical elasticity of matter can withstand them. When the nuclear energy sources are exhausted, the contraction of such a massive star occurs uncontrollably and irreversibly: it collapses and forms a black hole.

Everybody has heard about black holes. Everything about them strikes the imagination: the gravitation they create is so immense that light cannot escape them, and the rays of light passing in the vicinity of a black hole are curved and trapped by it. Even the geometrical properties of space and the course of time near black holes change in a most odd manner. Black holes produce bottomless funnels in space which "suck in" everything, from light to matter particles. The radius of such a funnel is comparable to that of a neutron star, amounting to several kilometres.

Much could be said about black holes. (The reader can find more detail about them in the books by I. D. Novikov (1976), S. A. Kaplan (1977), and I. S. Shklovsky (1977) listed in Recommended Literature.) But the most essential problem is whether they exist in nature. From the viewpoint of theory, black holes are just as unavoidable an outcome of the evolution of stars as white dwarfs or neutron stars. But it is clear that observations of black holes are hampered by their very nature. It would be hopeless to look for them as black points in the sky; their existence is rather to be deduced from indirect phenomena.

For instance, if a black hole and a common star compose a binary system, the peculiarities of the motion of the common star can help establish that its invisible companion is a black hole. Both stars in a binary system revolve about their common centre of mass, and measuring the orbital parameters of the common star, one can estimate the mass of its invisible companion. Naturally, a companion that is a weak but otherwise quite a common star would not be seen either. However, if the mass of the companion star is estimated at over two Sun masses, this suggests that the invisible partner is a black hole. If it were a common star of a large mass, its radiation would be registered. (Recall that both white dwarfs and neutron stars, which could be invisible in a binary system, possess smaller masses.) There are several sources in the sky suspected of being binary systems with black holes. For instance, such is a source of X-ray radiation in the Swan (Cygnus) constellation. The source is called Cyg X-1 for short. However, this is not a hundred per cent reliable, and new observations are subject to strict and thorough analysis. And still, one can hardly doubt the existence of black holes: massive stars do not have any other evolutionary path than the transformation into a black hole at the final stage. Most likely there are millions, if not hundreds of millions, of black holes in our Galaxy. And we must simply learn how to find them. Ya. B. Zeldovich once jokingly remarked that there are black holes wherever it has not been proved that they are not!

Close Binary Stars

The evolution of stars is a vast field of research in modern astrophysics, and our concise outline is far from being comprehensive. There are many popular-science books in the field, and an eager reader can become more closely acquainted with it. Besides the above-mentioned books by S. A. Kaplan (1977) and I. S. Shklovsky (1977), we can suggest the books by R. J. Tayler (1970, 1972) and Yu. N. Efremov (1980) listed in Recommended Literature. In concluding this chapter on the origin and evolution of stars, we would like to say a few words about new research which has hardly even been discussed in the literature for nonprofessionals. We are going to treat the physical processes in close binary systems of stars and discuss in more detail a remarkable type of these systems associated with a comparatively recently discovered phenomenon of cosmic bursts of X-ray radiation.

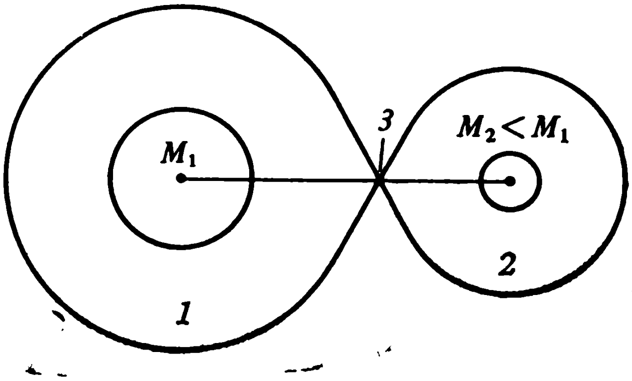

Fig. 40
A binary system of stars. *1*, *2*—the Roche lobes; *3*—the inner Lagrangian point.

If two stars form a close enough system such that the distance between them is comparable to their radii, the interaction between the partner stars is not limited to their revolving about the common centre of mass. It is important that in this case the matter of one star can flow to its counterpart under the effect of gravitational attraction.

Each star in a close binary system has its "zone of in fluence" within which its own gravitation, rather than the partner's, prevails. Such a zone is called the Roche lobe (E. A. Roche was a French physicist and mathematician of the 19th century who studied the mutual gravitation of planets and their satellites and developed a general theory applicable to binary stars, or simply binaries). These lobes must evidently contact each other at a point on the line connecting the stars' centres. The force of gravitation is zero at this point because each star here creates a gravitation identical to that of the other but opposite in its direction (Fig. 40). This point has a special name—the inner Lagrangian point (J. L. Lagrange was a famous French mathematician and astronomer of the late 18th-early 19th century). If the masses of the stars are the same, the inner Lagrangian point is equidistant from them; but if the masses are different, it is naturally closer to the less massive star because the Roche lobe is more extensive for the star of a greater mass. In this instance, the position of the inner Lagrangian point is determined by both the gravitation of the stars and the fact that this point revolves, as the stars do, about the common centre of mass of the binary system.

The matter exchange between stars can occur in two possible ways: either a "star wind" penetrates from the Roche lobe of one star into the Roche lobe of the companion star or one of the stars becomes greater than its Roche lobe.

Star wind was first discovered for the Sun (solar wind); it is a continuous supersonic flow of plasma from the solar corona into interplanetary space. If a star is hotter and more massive than the Sun, its flow of plasma is more intense; these star winds have sufficient enough velocities and kinetic energy resources to overcome the gravitation and thus leave the star for good. As to a binary system, some particles leaving one star can be trapped by the gravitational field of the other star.

Much greater portions of matter can flow from one star to the other in the second case, i.e. when one of the stars is greater than its Roche lobe: the flow of matter is not limited by the plasma emission from the coronas. The considerable transport of matter from one star to the other is capable of significantly influencing the nature of the further evolution of both stars in a close pair.

Many interesting things about the processes of this kind have been clarified fairly recently in the work carried out by A. G. Masevich, A. V. Tutukov, and L. R. Yungelson. The more massive star of the pair enters the stage of evolution when it sheds its envelope. A considerable portion of matter of this envelope can be captured by the second, less massive star, but since the mass of the star is thus increased, the latter star becomes more massive, and therefore the rate of its evolution speeds up. Before long, it begins to expand, and the size of its envelope becomes so vast that the remnant of the first star turns out to be within this envelope. The first star, which is now a neutron star, moves within the envelope of the larger star and its motion is impeded just as an artificial satellite slows down in the denser layers of the Earth's atmosphere, and thus it approaches the core of the second star, and finally, they form a close double core within a single big envelope. It is remarkable that such objects, i.e. two compact stars in a common envelope, have recently been found by direct astronomical observations.

Interesting events in a close binary system can also develop when the flow of matter from one star to the other is not too significant, but still one of the stars has turned into a neutron star. This is connected with the discovery and research on the burster sources of X-rays.

The fact that stars can emit both visible light and invisible electromagnetic waves of the X-ray range has become clear during the past quarter of a century when special X-ray telescopes appeared and were mounted on balloons, missiles, and artificial satellites and thus were carried out beyond the Earth's atmosphere, which prevents the penetration of cosmic X-rays. No less than a hundred such X-ray stars are known to belong to our Galaxy. The Sun is not included in the list; although its X-ray radiation has been registered since 1948, it is noticeable only because the Sun is close to us. The solar X-ray flow would be quite intangible if emitted from the distances at which the galactic sources of X-rays are located.

X-rays are electromagnetic waves belonging to the shorter wavelength range than both visible light and ultraviolet radiation. They correspond to wavelengths from 100 to 0.3 Å (1 Å (angstrom)$=10^{-10}$ m); hence X-ray photons have energies from 0.1 to 30 keV (kilo-electron volts), or 10^{-20}-3×10^{-18} J. Considering the X-ray range, the brightest X-ray stars emit ten thousand times more energy than the Sun does to all wavelengths.

Strong variability is a feature of X-ray stars. Some of them possess exceptional periodicity, i.e. strict regularity of brightness variations, with a period of seconds or fractions of a second. These are the X-ray pulsars which have already been mentioned. Bursters are a special class of X-ray variable stars producing short and very bright bursts of X-rays with a duration from several seconds to several minutes. The bursts follow each other irregularly, without exact periodicity. The energy of the X-rays in each of such bursts is as much as the Sun's energy emitted across the spectrum for weeks.

The history of bursters began in 1975, when a group of Soviet researchers reported short and intense bursts of X-ray radiation registered by the apparatuses mounted on the *Kosmos-428* satellite. Very soon, American astronomers detected bursts of X-ray radiation from the centre of the globular star cluster NGC 6624, where an X-ray star, one of the brightest sources in our Galaxy, had been discovered several years earlier. There are 32 bursters known to date, and eight of them apparently belong to the globular clusters of our Galaxy.

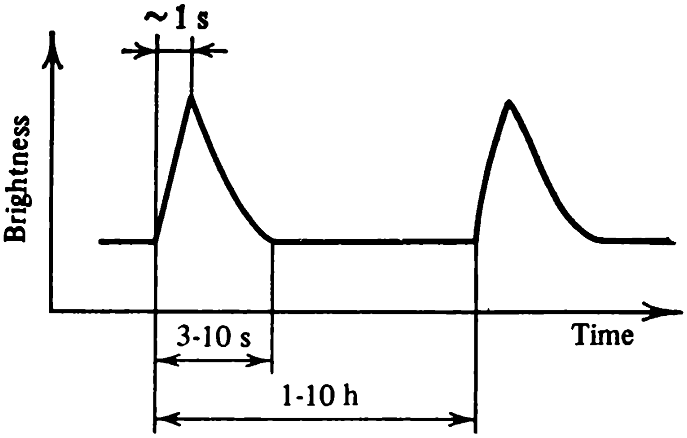

Fig. 41
A typical time dependence of a burster's brightness.

A sudden gain in the brightness of a typical burster usually occurs during a period from a few fractions of a second to ten seconds (Fig. 41). Then the brightness declines to the initial level during a time from several seconds to several minutes, and a new burst can be expected a few hours or days later. There are bursters that sometimes "switch off" for weeks and months on end, and then "revive" and give bursts of X-rays again.

Bursters flare up almost always against the background of steady X-ray luminosity of a star, although the time of this luminosity varies slightly.

The location of bursters in the sky is far from being random or uniform (Fig. 42). Most of them are concentrated in the direction to the centre of our Galaxy, and this became observational evidence that they belong to our

Galaxy. Another deduction is that the typical distance to bursters is of an order of 10 kpc; this is a distance comparable to that between the Sun and the galactic centre.

Measuring the flow of X-rays caught by the telescopes and knowing the typical distance to the source, one can find the proper luminosity of the source, i.e. the total energy emitted by it per second. Such calculations proved that a burster emits from 3×10^{30} to 3×10^{31} J per second during a burst. (Recall that the total luminosity of the Sun is 4×10^{26} W.) The total energy of the burst, i.e. the luminosity times the duration of radiation, amounts to 3×10^{31}-3×10^{32} J.

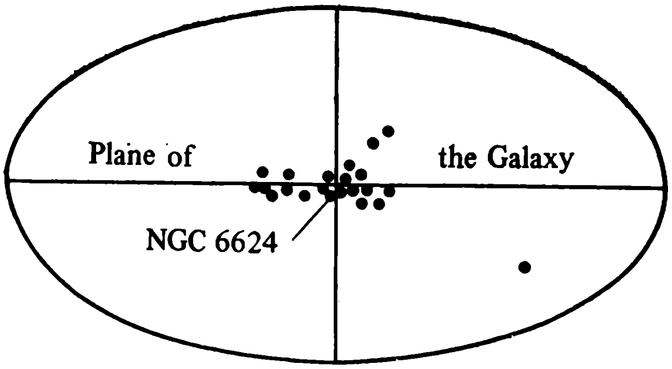

Fig. 42
The localization of bursters in the Galaxy.

The proportion between the burst luminosity and the steady background luminosity of a burster is its important characteristic. It was found that tens and hundreds of times more energy is emitted during the time between two typical bursts, i.e. during the quiet period, than in each individual burst, however bright it might be.

Another feature of bursters is no less important for understanding their nature. It has been established, proceeding from the pattern of the spectrum of their X-ray radiation, that during its flare a burster is as luminous as the surface of a body heated to a temperature of about 30 million kelvins. At such a temperature, any body emits most of its energy in the X-ray range of electromagnetic waves.

Knowing the luminosity patterns of heated bodies and the value of emitted energy per unit time, one can calculate the surface maintaining a given luminosity at a given temperature. It was estimated that during its flare a burster emits approximately the same amount of energy as the surface of a sphere of radius of about 10 km would emit when heated to 30 million kelvins. But this is the size of the most dense cosmic bodies, i.e. neutron stars.

The conclusion that X-ray bursts are associated with neutron stars was not unexpected. Similar objects, X-ray pulsars, were known at that time, and there is no doubt that they are neutron stars. There is every reason to think that the radiation of X-ray pulsars is related to the process of accretion, i.e. the capture of external matter by the gravitational field of a neutron star. Ya. B. Zeldovich, I. D. Novikov, and I. S. Shklovsky revealed that this phenomenon is the source of energy in all X-ray stars.

The captured matter is drawn or flows from the surface of the companion star in a close binary system, and then it is accelerated in its free fall in the gravitational field of a neutron star and acquires a considerable velocity and kinetic energy. The unavoidable collision with the surface of the neutron star causes the conversion of the kinetic energy of the falling matter into heat. The result is that the surface of the neutron star, be it the entire surface or a spot on it, is heated to temperatures of millions and tens of millions of kelvins and begins to radiate, emitting mainly X-rays (in keeping with the temperature values).

The accretion in X-ray pulsars is controlled by the magnetic field; the field prevents matter from moving across the lines of force, and therefore the matter falls on the surface of a neutron star only through magnetic field "funnels" around the magnetic poles rather than uniformly from all directions. (It is well known that there are such funnels for the interplanetary matter and cosmic rays near the magnetic poles of the Earth: this is why polar auroras can be admired there.) It can be surmised that this kind of directed accretion results in hot spots on the surface near the poles of a neutron star (Fig. 43). The rotation of a neutron star about its axis is inclined (as is the Earth's) with respect to the magnetic axis passing through the magnetic poles, and this creates the beacon effect: the bright spot is now visible and now invisible. The period of observation of the spot is the period of rotation of a neutron star: this is why it is so highly regular.

It is natural to assume that bursters are also neutron stars in close binary systems rather than single neutron stars, and there is a flow of matter from one star in the system to the other and therefore accretion. But how come in some cases a neutron star is an X-ray pulsar and in other cases it is an X-ray burster?

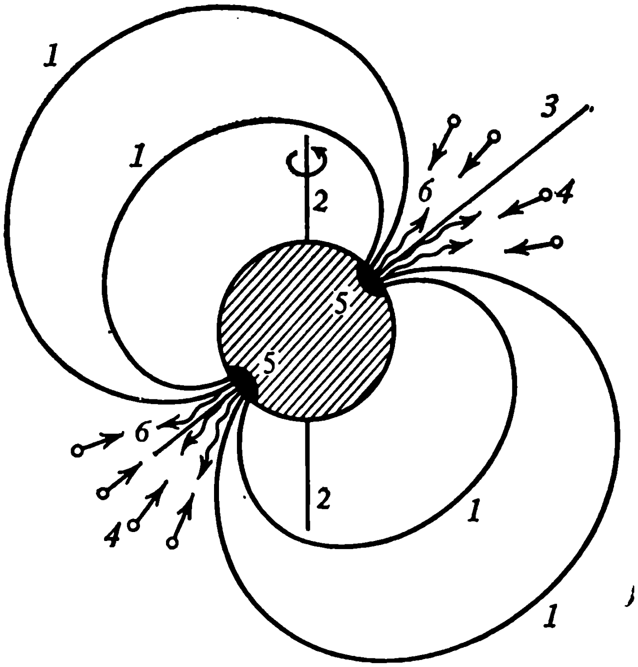

Fig. 43
A diagram of an X-ray pulsar. The accretion of matter by a magnetized star. Radiation from hot spots. *1*—lines of magnetic force; *2*—the rotation axis of a neutron star; *3*—its magnetic axis; *4*—fall of particles into magnetic funnels; *5*—hot spots in the vicinity of the magnetic poles; *6*—radiation from the spots.

The fact that the radiation pulses from bursters follow each other without any regular period evidently suggests that the mechanism of bursters is related neither to the proper rotation of a neutron star about its axis nor to its periodic orbital revolution in the binary star system. Otherwise bursters would be just as good timers as pulsars are.

Continuing this discussion, we apparently have to admit that, in contrast to an X-ray pulsar, there should be no hot bright spots on the neutron star of a burster. This implies that the magnetic field creating funnels and hot spots under them in pulsars is either absent or not strong enough to control the flow of matter from the "normal" star to the neutron star in bursters.

When the accretion is not controlled by the magnetic field, the fall of matter is more or less uniform over the entire surface of a neutron star. Then the whole surface can be heated to a high temperature, and can emit X-rays during the continuing process of accretion (Fig. 44). It seems reasonable to relate the radiation of this origin to the steady background X-ray flow which, as has been mentioned above, is registered from most bursters.

Observations give an estimate of an order of 10^{30} W for the background luminosity of bursters. It is easy to evaluate the rate of accretion capable of maintaining such luminosity. While matter falls in the gravitational field of a neutron star, it is accelerated to the velocities close to the velocity of light, i.e. 3×10^{10} cm/s. (More accurately, these velocities are from a fifth to a third of the velocity of light.) The corresponding kinetic energy converts into heat when the matter collides with the surface of a neutron star, and then the energy is emitted as a flow of X-rays. The exhibited luminosity is half the square of the velocity of the fall times the mass of matter reaching the surface of a neutron star per second. It is easy to show that the observational value of the background luminosity indicated above can be maintained if 10^{17} grams of matter fall on the surface of a star each second.

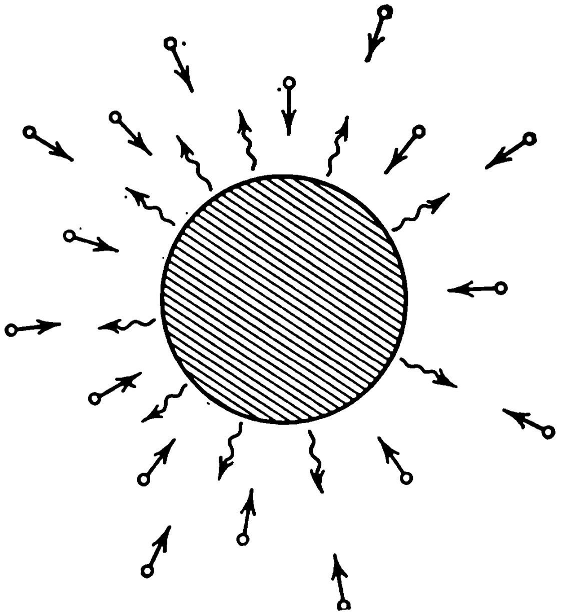

Fig. 44
A diagram of 'a burster. The accretion, heating, and radiation of the entire surface.

It is interesting that the same value of the flow of accreted matter is known independently from the data on X-ray pulsars. The coincidence is hardly accidental: it is rather an indication that the nature of the background luminosity of bursters discussed above is in agreement with the general picture of the physical processes in close binary systems.

What is the fate of the matter falling on the surface of a neutron star? Fresh gas captured by a neutron star is rich in hydrogen. This is the matter lost by the outer layers of the "normal" star, the companion of the neutron star in the close binary system, and it contains no less than 70-75 mass per cent of hydrogen, about 25-30 per cent of helium, and about one per cent of heavier elements. Thermonuclear reactions of transformation of hydrogen into helium can occur when hydrogen is heated to high temperatures and intensely compressed to high densities upon collision with the surface of a neutron star. These reactions are the same as within the Sun and similar stars.

The released energy of thermonuclear reactions enhances the overall heating of the surface of a neutron star. This complement is small compared with the kinetic energy of the matter fall. However, it is more important that thermonuclear reactions on the surface of a neutron star produce a layer of helium. And it is precisely here that the characteristic burster processes develop.

According to calculations, this layer is about one metre wide, its density is over a million g/cm^3, and its temperature is about 500 million kelvins. The reaction of fusion of three helium nuclei into a carbon nucleus can occur under such conditions; nuclear physicists call this reaction the triple-α process (a helium nucleus is an alpha particle). The triple-α process is extremely sensitive to temperature and therefore can develop like an explosion. It occurs so rapidly in the helium layer that its temperature, for instance, can jump, doubling in a few hundredths of a second. Once started, helium burns at an ever increasing rate until all of it becomes carbon.

The possibility of an explosive thermonuclear burning in the outer stellar layers was proved theoretically by L. E. Gurevich and A. I. Lebedinsky over three decades ago. Detailed calculations for bursters were carried out by V. Lewin, P. Joss, and E. V. Ergma.

Thermonuclear reactions in a gram of helium release 10^{11} J of energy. How much energy is released during a burst? To answer this question, it is evidently necessary to know the mass of the helium layer.

Most probably, helium completely transforms into carbon during a single explosion; it is accumulated during the time between the explosions. To make an estimate, take the interval between two X-ray bursts of an order of 10^4 s, and use the estimate of the rate of accretion obtained above (10^{17} g/s); the result is that the mass of the layer amounts to 10^{21} g. Now multiply this value (by the way, it can also be deduced from the parameters of the helium layer indicated above) by the energy released while a gram of helium converts into carbon, and therefore estimate the total energy of the burst: 10^{32} J. This value is in close agreement with observational data, which has already been mentioned.

We can also estimate the typical duration of a burst. It is only necessary to make use of the data of nuclear physics on the rate of energy release in the triple-α process as it depends on the density and temperature of matter. Using the characteristics of the helium layer indicated above, we get about 10^{10} J of energy released per gram of matter per second. If the mass of the layer is 10^{21} g, the total rate of energy release is about 10^{31} W. This quantity is very close to the typically observed luminosity of a burster during an X-ray flare.

Theory also well explains the ratio between the peak and the background luminosity of a burster. This makes it possible to believe that the nature of burster radiation has been revealed.

We have made use of bursters, a fairly recent and interesting discovery, in order to discuss, in more detail than in other cases, the astronomical observations and the physical theory leading to the solution of some puzzling problems and then to the creation of still other problems in the study of the infinitely diverse stellar universe. The life of stars from their birth to the final stages of their evolution, from protostars revealing themselves as interstellar masers to neutron stars, pulsars and bursters, becomes a clear-cut chain of amazing transformations, a sequence of explosions, flares, and bursts.... Today we can hope to be closer "to understanding, at last, such a simple thing as a star", as A. S. Eddington once remarked.

Chapter 5

The Evolution of Stellar Systems

The history of each galaxy includes a short but eventful epoch when its matter, i.e. a cloud of gas separated from a protocluster shortly before, was compressed by its own gravitation. This process produced the first stars and shaped galactic subsystems such as the halo and the disk of our Galaxy. The cloud of gas transformed into a stellar system, and the stellar system attained its steady, stationary state. Further changes in the system were of quite a different nature, and they took place much more slowly. Fast and active processes only occurred in the central areas of galaxies, i.e. the galactic nuclei. Furthermore, spiral galaxies continued and still continue to form young stars out of gas and dust from their flat subsystems where the spiral pattern outlined by these stars is observed.

This chapter deals with the evolution of stellar systems. We shall discuss both galaxies and star clusters within galaxies.

From a Protogalaxy to a Stellar System

The separation of protogalactic condensations appeared to occur owing to hydrodynamic processes in gaseous protoclusters; the further evolution of protogalaxies was primarily controlled by their own gravitation, which contracted these thin clouds, thus shaping the galaxies of the observed sizes.

Gravitational contraction was unimpeded by the forces of pressure because the protogalactic gas could easily cool to a temperature of about ten thousand kelvins. The pressure corresponding to this temperature is unable to counteract the general gravitation of the entire protogalaxy. The density of a protogalaxy increased from 10^{-27} g/cm^3, which is characteristic for both protoclusters and present-day clusters of galaxies, to the typical galactic density of 10^{-24} g/cm^3. This increase happens to be reached in such a way that all the particles of the cloud fall freely in their common gravitational field. A thousandfold increase in the cloud's density corresponds to a tenfold decrease in its size.

How long does this free fall continue? It turns out that it takes approximately the same time as the period from the beginning of the cosmological expansion to the separation of protoclusters, i.e. about 3000 million years. This coincidence is hardly accidental. The point is that the dynamics of the cosmological expansion, first impeded and then halted owing to the development of the gravitational instability in the protocluster volume, is controlled by the forces of gravitation of matter and only by them. Just as well, the process of the contraction of a protogalactic cloud, i.e. a portion of a protocluster, is only controlled by the forces of its own gravitation. Consequently, the dynamics of a gravitating medium possesses a symmetry with respect to the "halt" moment when the expansion of given volume is replaced by its contraction. For instance, a tenfold expansion takes the same time as a tenfold contraction.

Directly before and immediately after the halt, the rates of expansion and then of contraction are small, and this "suspension" takes a lot of time. The further contraction occurs with acceleration (as in any free fall) and proceeds at a faster rate. Symmetrically, the expansion preceding contraction takes place first at a fast rate and then slows down. Thus the increase in size from a tenth of the final volume to its maximum requires no less than 95 per cent of the whole expansion time. Therefore the time it takes the galaxy to fall back from its maximum volume to a tenth of this size must be within several per cent of the period for the galaxy to begin its expansion to the time the expansion halts.

The contraction of a protogalaxy lasts about 3000 million years. It takes this time to transform a gaseous cloud into a stellar system, which no longer contracts and continues to exist in a steady, almost invariable state. The transition from contraction to a steady-stationary state is associated with the separation of the protogalaxy into fragments, within which the process of star formation starts. The gravitational instability developing in a protogalactic cloud produces condensations of 3×10^7-3×10^9 Sun masses. As we mentioned in Chapter 4, the masses of major fragments of a protogalaxy are determined under the conditions of gravitational instability by the Jeans criterion. Their size is estimated proceeding from the fact that the temperature of the protogalactic gas is close to ten thousand kelvins, while its density varies within 10^{-27} and 10^{-24} g/cm^3. It is most probable that a few hundred million years after the beginning of contraction, a protogalaxy becomes a combination of separate fragments, or denser clouds, rather than a continuous cloud of gas. Consider a protogalaxy with a mass of 100,000 million Sun masses, i.e. comparable to the mass of our Galaxy. There may be several dozen major clouds with masses up to a thousand million Sun masses and hundreds or thousands of smaller clouds with masses of a few tens or hundreds of millions of Sun masses.

All these numerous clouds fall freely to the centre of a protogalaxy. However, in addition to their free fall velocity, the clouds have their proper motions which were produced, as the clouds themselves, by the gravitational instability within a protogalaxy. The chaotic, irregular proper motions of the clouds make them collide with each other, and therefore the material of the clouds is heated and compressed owing to such collisions; some of the kinetic energy of the random cloud motion is spent during this process. Cooling follows, and radiation withdraws this energy from a protogalaxy. This energy release makes possible further general contraction of the whole combination of clouds.

It is essential that since the clouds become denser owing to the collisions with each other, cascade fragmentation rapidly develops in them and results in the formation of the first stars in a galaxy. The most massive stars appearing in the process have enough time to progress through the entire cycle of their evolution from the state of a protostar to the Sun-like state and then, having exhausted their nuclear energy resources, they may explode as supernovae. A star of a few dozen Sun masses requires for this no more than a few dozen million years, which is noticeably less than the time of the general protogalactic contraction.

The explosions of supernovae make protogalaxies rich in the products of thermonuclear synthesis within massive stars, i.e. in elements heavier than hydrogen and helium. Consequently, as a result, new groups of stars are created out of matter more and more abundant in carbon, oxygen, nitrogen, and metals. If a given group of stars appears later, the hydrogen-helium interstellar medium is more abundant in heavy elements, and the stars produced out of this medium are richer in these elements. Since during this process the gaseous protogalaxy continues to contract, the star formation occurs closer and closer to the centre of the given group of stars and gaseous clouds. Because of this process, it can be expected that the stars of the inner areas of galaxies should differ in their composition from the stars of the outer areas: the closer to the centre, the greater the abundance of heavy elements in the stars.

This conclusion is in close agreement with the astronomical data on the abundance of heavy elements in the stars of the halo, i.e. the spherical subsystem of our Galaxy, as well as in the elliptical galaxies similar in their structure and star composition to the halo of our Galaxy. This confirms the general concept of gaseous protogalaxies and the formation of stars during their gravitational contraction.

Until now, we have not taken into account the fact that a protogalaxy may possess considerable rotation. In fact, rotation is not quite significant at the initial stages of contraction, while inertia is weaker than gravitation. But rotation should become faster owing to contraction. This concerns both a protogalaxy and, in general, every body whose size decreases. At the same time inertia increases. Inertia does not impede the contraction along the rotation axis, but it is capable of counteracting the force of gravitation in the directions at right angles to the axis. This is why a contracting cloud, almost spherical at the beginning, becomes more and more flattened and tends to take the shape of a disk in which the contraction along the directions transverse to the rotation axis first slows down and then halts completely when inertia balances gravitation in these directions.

This general tendency can be revealed in a contracting protogalaxy. But it has to be borne in mind that part of its matter turns into stars at the initial stage when rotation does not yet influence the general shape of the protogalaxy. Such stars are the ones of the halo, i.e. the spherical subsystem of our Galaxy. The remaining gas continues to fall, and while it becomes denser, its rotation becomes faster until inertia halts the contraction at right angles to the rotation axis; the contraction along the axis con-

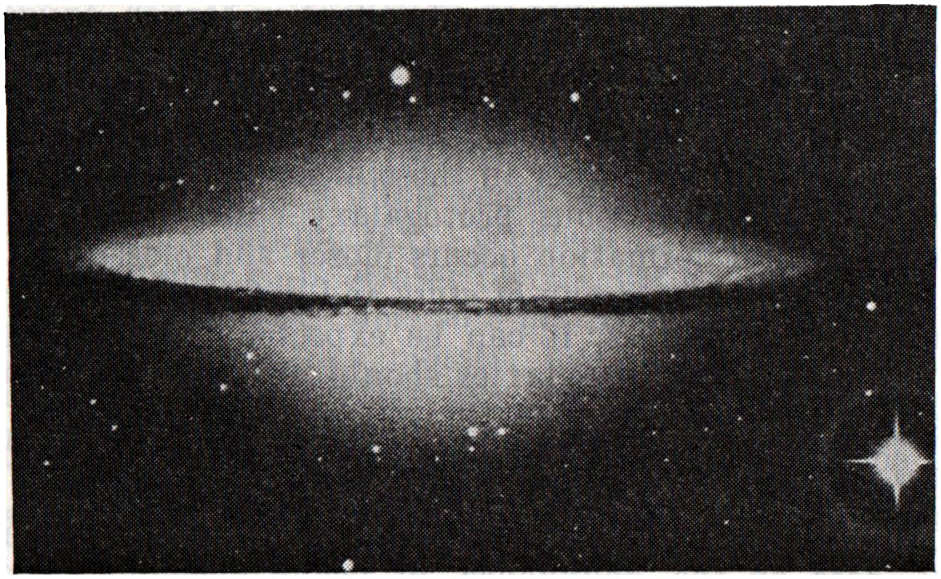

Fig. 45
A spiral galaxy with a layer of dark gas-dust clouds in its disk.

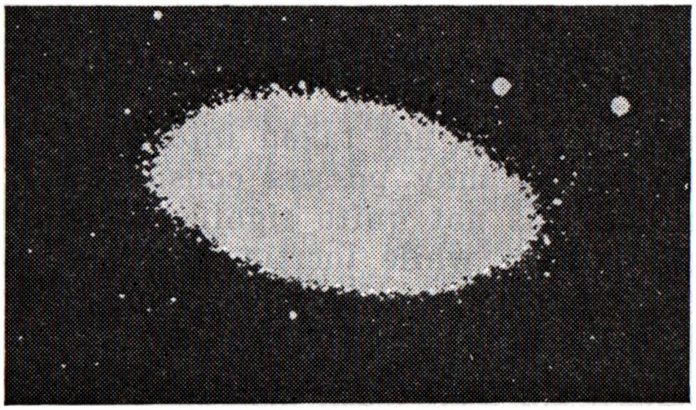

Fig. 46
A small elliptical galaxy, a satellite galaxy of the Andromeda galaxy.

tinues and at long last results in the formation of a rotating disk, i.e. the flat subsystem of the Galaxy. Other spiral galaxies with fast rotation of their flat subsystems have been formed in much the same manner (Fig. 45).

If the rotation of a protogalaxy from the very beginning is either very weak or absent completely, there appears to be no mechanism leading to the formation of a disk in it. This is the case when the general contraction of a protogalaxy and the formation of stars in it result in the shaping of a more or less spherical system. This is in fact what produced elliptical galaxies devoid of any fast rotation (Fig. 46).

The Motion of Stars in Galaxies

An elliptical galaxy is a composite system of stars related by their common gravitation. Each star moves in the common field of gravitation of the system and is not affected by any forces besides gravitation. It can be said that each star falls freely in the gravitational field of the system. Falling from a height to the centre of the system along the radius, the star is accelerated, and when it reaches the central area, it overshoots it and then begins to withdraw from it towards the opposite part of the system. Now the motion is decelerated rather than accelerated (like the motion of a stone thrown vertically upwards), since this motion is oriented against the force of gravitation, which is directed, as always, to the centre. As the star moves from the centre, its velocity decreases, and finally the star becomes "suspended" for an instant (as the stone at the highest point), and its velocity vanishes. Then it begins another fall to the centre, the star reaches and overshoots the centre again, and returns again to the position from which it began its fall. This cycle repeats itself over and over.

There are no other forces besides gravitation in a stellar system, but nonetheless the system does not contract but rather remains stable as a whole. This is possible because each of the stars in the galaxy moves periodically around an orbit that is confined to a volume at the borders of which the velocity of the star's movement away from the system's centre vanishes.

It is not at all necessary that the orbit of each star pass exactly through the centre of the system. Stellar orbits are different, and besides purely radial orbits, elliptical orbits are also possible, similar to the orbits of the planets in the solar system. However, stellar orbits are less circular, i.e. much more elongated. All these orbits uniformly fill the volume of a stellar system, producing its regular spherical or more or less ellipsoidal, elongated shape. This is how elliptical galaxies and the spherical subsystems of spiral galaxies are arranged.

The disks of spiral galaxies and, in particular, the disk of our Galaxy have a lot in common with the solar system in terms of their dynamics. Like planets, every star of the disk moves around an almost circular orbit, where inertia is balanced by gravitation. But unlike the solar system, the force of gravitation is created in galaxies primarily by the stars of the disk and the spherical component of the galaxies rather than by a central body (galactic nuclei are not very massive).

Average velocities of stellar motions in our Galaxy, both around elongated and circular orbits, amount to 100-300 km/s. The velocities are small in less massive galaxies and greater in more massive galaxies, but they are always between tens and a thousand kilometres per second.

Violent Relaxation

The shape and internal structure of elliptical galaxies or spherical subsystems of spiral galaxies are too regular and harmonious to be entirely explainable by the initial shape and structure of protogalactic clouds. Most probable, protogalactic clouds were loose and patchy and had no distinct and regular boundaries. So what gave stellar systems their regular shape and structure?

Naturally, this is also a result of the action of gravitational forces: there are no other forces in stellar systems. It has long since been known that the forces of gravitation always tend to give a regular rounded shape to celestial bodies; this is evidenced by the shape of the planets, the Sun, and other stars. The tendency to such a regular shape and structure is similar to the one exhibited during the process of relaxation in other physical systems, for instance, such as a gas of atoms or molecules.

Relaxation is a process in which a physical system approaches a steady, equilibrium state. The relaxation in a gas results in a steady state of general uniformity in the distribution of atoms or molecules in the volume they fill; it also leads to the equilibrium distribution of random thermal velocities of particles. In thermodynamics, the equilibrium velocity distribution of particles is called the Maxwell distribution. The relaxation in a gas is carried out through collisions of particles with each other; random collisions of particles change their velocities, and thus the Maxwell distribution is reached.

Stellar systems reveal certain signs of an equilibrium state. Besides their regular shape, this is indicated by the velocity distribution of stars, which resembles the Maxwell distribution; it can be confirmed by the data on the stars in the vicinity of the Sun in our Galaxy. But collisions of stars in galaxies, similar to collisions of gas particles, are impossible. Stars do not collide head-on and, in fact, do not approach each other close enough to change their velocities even in the least because of pair gravitational interaction. Estimates show that the expected time of a close approach between two stars is much greater than the age of galaxies. This is why it can be said that galaxies are collisionless systems.

And nonetheless, the relaxation in stellar systems is possible. In 1967, D. Lynden-Bell put forward the idea that relaxation is due to the interaction of each star with the gravitational field of the whole system rather than to pair "collisions" of stars. This collisionless relaxation took place during the epoch when galaxies were just forming. During the general contraction and fragmentation of a protogalaxy, the gravitational field of the system greatly varied with time and from site to site. Therefore the stars born in the process underwent rapid changes in their velocities and orbits. The ensuing changes in stellar motions were rather strong: it can be said that an individual star collided with the whole system or at least with its major portions.

This process was random in the same way in which collisions of gas particles were random. The results of each individual change in the motion of a given star were unpredictable, but there were many such changes and they occurred continuously, and therefore their total action made possible the tendency to reach an equilibrium (this tendency was common for all physical systems).

All this gives grounds to see the features of relaxation in such a process. D. Lynden-Bell called it "violent relaxation". The relaxation was actually violent: it occurred in an unsteady, strongly "excited" state of a protogalaxy when there were both the first stars and the chaotically moving (with velocities of hundreds of kilometres per second) massive gaseous fragments.

The theory of violent relaxation, as well as the entire complex of problems related to the transformation of a protogalaxy into a stellar system, still remains insufficiently elaborated. However, scientists succeeded in finding out many things with the aid of modern computer simulation rather than by means of theoretical calculations. It is possible to programme a computer to find the time dependence between the velocity and position of a star in a system, and thus clarify the general behaviour of the system as a whole. It turned out that a collisionless system of gravitating bodies actually tends to arrange itself as a regular spherical formation to become more flattened with time. This occurs during a period comparable to the typical period of revolution of a star in a system. It is noteworthy that the result is valid for widely diverse initial states, which can serve as the beginning of any computer simulation.

The Evolution of Star Clusters

In contrast to galaxies, globular and open (galactic) star clusters are not collisionless systems. (Star clusters are treated in the book by Yu. N. Efremov (1980) listed in Recommended Literature.) Owing to the greater density of stars in star clusters, each star experiences at least several close encounters and many distant encounters with other stars during the period a cluster exists. The time between such "collisions" is still greater than the period of revolution of stars around their orbits. The sizes of the orbits are comparable to those of the clusters, so that during one revolution there is only a low probability for a star to encounter any other star. It can be said that the mean free path of a star, i.e. the path between two encounters with other stars, is considerably greater than the size of the whole system. These circumstances determine the nature of evolution of star clusters.

The relaxation in star clusters is possible owing to pair encounters between stars. These systems tend to the state of equilibrium, and gravitating systems in general are as close to this state as is possible. The resulting velocity distribution of stars is close to the Maxwell distribution mentioned above.

But star encounters sometimes lead to the following: a star can accidentally acquire such a great velocity that it overcomes the attraction of other stars and leaves the system. Therefore, the velocity distribution of stars, in contrast to that of gas particles, has no stars in the high-speed "tail" of the velocity distribution. Fast stars account for about a hundredth of the total number of stars in the system. This inevitable and steady deviation from the state of equilibrium associated with a constant "evaporation" of stars is a characteristic feature of a gravitating system.

During their "evaporation" evolution, star clusters become more and more nonuniform in their density, and each of them develops a compact central area, i.e. a nucleus, surrounded by a comparatively rarefied halo (Fig. 47). Losing the fastest stars, a cluster can finally disintegrate completely and diffuse, as was first shown by V. A. Ambartsumyan in 1938.

The general theory of dynamic evolution of stellar systems, including the phenomena of relaxation and "evaporation" of the fastest stars, was developed during the 1940's-1950's by S. Chandrasekhar, K. F. Ogorodnikov, L. E. Gurevich, B. Yu. Levin, and T. A. Agekyan. An important result is that many star clusters or groups originally produced in a galaxy might become almost completely destroyed by the "evaporation" evolution, and their stars could diffuse and scatter over the whole volume of the system, which is observed in real galaxies.

It appears that the fates of dense clusters of stars in the central areas of galaxies are different. Because the fastest stars "evaporate" from these star clusters, they generally contract. Pair encounters of stars happen in them more and more often, and stars in such cases approach each other more and more closely. Finally, at a certain stage of the evolution of the system, direct collisions between stars become possible.

Such a collision is a direct contact between two stars, and it is clear that their internal structure can be strongly altered: the stars can undergo deformation, split into parts or, on the contrary, stick to each other. It is most probable that the stars' outer layers are "stripped off"; the released gas first scatters over the system and then precipitates to its centre, undergoing gravitational condensation and fragmentation there. This makes possible the formation of young stars associated into a dense, bright, and condensed subsystem. There are certain conditions (for instance, when the gas temperature is very high), at which a single supermassive star rather than a stellar subsystem is formed.

Fig. 47
A globular cluster in the Toucan (Tucana) constellation.

If the original star cluster was massive enough and contained, for instance, 10^9 to 10^{10} stars, then the supermassive star could possess a mass of 10^6-10^9 Sun masses. Very intense radiation is the main feature of such a star.

For instance, if its mass is 10^8 Sun masses, the luminosity of the star amounts to 10^{39} W, so that the supermassive star, which comes into being, if possible, in the dense central area of such a stellar system as a major elliptical galaxy, can increase the luminosity of the system as a whole by a factor of ten to several hundred times. (Recall that the luminosity of our Galaxy is about 3×10^{37} W.)

A supermassive star emits light at the expense of its potential gravitational energy and gradually contracts. A collapse occurs at the final stage of contraction, i.e. an uncontrollable fall of matter to the centre, which cannot be counteracted by any pressure, and a black hole appears. However, the central area of the system can emit energy long after that, but this emission is produced by the gas and stars accelerating to immense velocities and colliding with each other while falling into such a supermassive black hole.

The evolution of stellar systems at the stage of contact collisions includes a number of other important processes. Thus, if the collisions between stars are nonelastic, they may bring about a coalescence of stars. The evolution of more massive stars thus produced is faster, and S. Colgate suggested that there can be numerous outbursts of supernovae in such systems during a period of about 10^6 years. This also results in a considerable increase in the luminosity of the system. After the burst, such a star becomes a neutron star (e.g. a pulsar), and if it is a very massive star, it may collapse, producing a black hole. The luminosity of the system is thus complemented by the radiation of pulsars and the flow of energy due to the accretion of matter by black holes.

The outward features of the central areas of condensed stellar systems at the stage of contact collisions resemble those of quasars or the active nuclei of galaxies. (The reader can find more detail in the books by E. A. Dibai (1977) and F. Hoyle (1965) listed in Recommended Literature.) However, we cannot treat this problem thoroughly here, although it is currently the subject of numerous investigations. It is extremely complicated and is yet very far from its solution: it is so important that a separate book should be written on it alone.

Chapter 6

"Hidden Masses", Neutrinos, and Einstein's Vacuum

The existence of unseen but essential masses of matter in the universe was first suspected some 50 years ago, when astronomers started to investigate groups and clusters of galaxies. During recent years, observations helped discover that both our Galaxy and other major galaxies possess vast coronas extending far beyond the visible stellar systems. Coronas do not emit light, but their mass is believed to exceed the total mass of the stars they surround. The nature of this "hidden mass" is a most perplexing puzzle in astronomy. However, recent discoveries in the physics of elementary particles give hope for the solution of the problem: galactic coronas seem to be filled with neutrinos, tiny particles of matter.

Astrophysics and the physics of elementary particles are two branches of science which study natural phenomena of the greatest and the least characteristic lengths. They exhibit an ever growing tendency to reveal a profound internal relationship between these two space scales. It becomes increasingly evident that the properties of elementary particles considerably control the structure and evolution of stars, galaxies, and the universe as a whole, while the space distribution of elementary particles and, probably, their very origin are associated with the violent processes in the early universe.

The most important ideas, hypotheses, and opinions currently discussed in both cosmology and the physics of elementary particles in connection with the picture of the early stages of the universe are covered in the papers by Ya. B. Zeldovich (1977, 1981), A. D. Dolgov and Ya. B. Zeldovich (1980), and the book by S. Weinberg (1977) listed in Recommended Literature. We are going to treat this complex of problems in order to consider large-scale astronomical phenomena occurring in the present-day universe that were caused by the physical processes at the level of elementary particles during the first instants of the cosmological expansion.

It turns out that many fundamental deductions in astrophysics and cosmology depend on the number of neutrinos in the universe and whether these elementary particles possess any rest mass. It is probable that neutrinos are the major contributors to the density of matter in the present-day universe and therefore define the dynamics of the cosmological expansion, the geometry of the universe, and the further fate of the entire universe. The role of neutrinos in both cosmology and cosmogony is essential even if the neutrino rest mass is a very small fraction of the electron rest mass.

"Hidden Masses"

One of the best studied clusters of galaxies is the Coma cluster (A1656) in the Berenice's Hair (Coma) constellation. While the constellation consists of some nearby stars of our Galaxy, the Coma cluster of galaxies is much farther, far beyond the Galaxy. The distance to it is about 140 Mpc, i.e. 4×10^{26} cm.

The distance to the Coma cluster of galaxies and other clusters and superclusters of galaxies is inferred from the red shift in the spectra of the constituent galaxies. The red shift is a result of the cosmological expansion, the general recession of these systems. The velocity of the recession of the Coma cluster from us, deduced from the observed red shift, can be estimated accurately enough at 6850 km/s. The Hubble law, which associates the velocity v of the mutual recession of systems with the distance L between them, $v = HL$, makes it possible to find this distance. The distance to the Coma cluster indicated above was calculated assuming that the Hubble constant $H = 50$ km/(s·Mpc).

If the distance to a cluster is known, we can determine its size. It is only necessary to measure the angle at which the cluster is observed and then use the simple trigonometrical relationships between the sides and angles in the triangle whose two vertices are at the periphery of the cluster and the third vertex is on the Earth (Fig. 48). The Coma cluster of galaxies is observed at an angle of about 100 minutes, or 1.7 degrees, or 0.06 radian. Now if the distance is 140 Mpc, the diameter of the cluster is about 8 Mpc, so the radius is 4 Mpc, or 10^{25} cm.

The velocity of recession of the Coma cluster away from us corresponds to the velocity of its motion as a whole, or, in other words, to the motion of the centre of mass of the cluster. But each galaxy within the cluster revolves about the centre, and these proper motions of galaxies can also be found from the red shifts in their spectra. Most accurately, this is how radial velocities are deduced, i.e. the projections of galactic velocities on the line of vision. These velocities are usually about the same within a cluster of galaxies, as it is observed in the Coma cluster. For each galaxy, the red shift differs slightly from the average value (the latter is naturally ascribed to the cluster as a whole). The typical velocities of the proper motions of galaxies in the Coma cluster are estimated to be about two thousand kilometres per second, i.e. 2×10^8 cm/s.

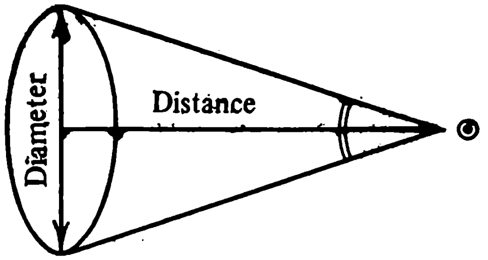

Fig. 48
Measuring the size of a cluster of galaxies.

There is a universal relationship pertaining to a steady-state gravitating system: it associates its mass M and its size R with the velocities v of the constituent bodies: $v^2 \approx GM/R$ (G is the gravitational constant). This relationship is called the virial theorem. It permits us to calculate the mass of a cluster by the known values of the proper velocities of galaxies and the radius of the cluster. This yields about 3×10^{15} Sun masses for the Coma cluster, which is approximately 10,000 times the mass of our Galaxy.

Such dynamic estimates of the masses of clusters of galaxies were made by F. Zwicky during the 1930's, and later by other astronomers.

Making such estimates, one has to be confident that the galaxies in a cluster are related by their common gravitation and move in a finite volume never escaping it. Astronomers were not at all completely and unanimously convinced of this when the study of clusters of galaxies was just started. Doubts were cast by the fact that the dynamic estimate of cluster masses contradicts another estimate based on measuring the brightness of galaxies within the clusters. This latter method proceeds from the fact that the ratio of the mass of a galaxy to its luminosity, i.e. the energy emitted per unit time, is identical for the galaxies of each given type. If this is so, then knowing the distance to the galaxies and their type, i.e. whether they are spiral, elliptical, or irregular, one can measure the flow of radiation from each galaxy and then calculate both the mass of each galaxy and the mass of the cluster as a whole. (In the final analysis, this method naturally proceeds from dynamics: when the standard ratio of the mass to the luminosity is found, the mass of a "model" galaxy is used, which is calculated by the motions of its stars.) The masses estimated in terms of the luminosities prove to be less than those based on dynamic estimates by an order of magnitude or more.

The alternative is either that clusters are unrelated galaxies and the dynamic estimate of the cluster masses is simply impossible or that the clusters contain considerable masses of matter which do not reveal themselves by radiation.

Consider the first case.

If a cluster is not steady, and no other mass besides the masses of galaxies is present within it, then the motion of the galaxies in the cluster characterizes the degree of instability of the system rather than its total mass. This viewpoint was offered by V. A. Ambartsumyan during the 1950's; it served as the central point of the cosmogonical hypothesis developed at the Byurakan astrophysical observatory.

V. A. Ambartsumyan suggested that clusters of galaxies appear following "explosions" of solid bodies; each such body is capable of splitting into fast receding fragments which later become galaxies. The energy of such an "explosion" imparts great velocities to the galaxies. The formation of stars is due to further separation of the fragments; some of the matter in its primordial state may continue to exist in the centre of galaxy. The prestar body in the nucleus of a galaxy can proceed through a series of immense explosions accompanied by ejections of considerable masses of matter, the emission of intense radio radiation, etc.

This cosmogonical concept does not give any concrete details on the nature of the prestar body. It simply makes an assumption that these bodies do not necessarily follow the known laws of physics; for instance, the conservation of energy can be violated during the explosions, and the conservation of momentum can be violated as well.

It is difficult to develop a theory if the hypothesis is based on such radical assumptions. By the way, during the 1920's, J. Jeans remarked, without going into detail, that the problems of explaining the origin of the spiral arms of galaxies create a suspicion that the centres of the galaxies are special points through which matter. may penetrate into our universe from other worlds. Indeed, this is also a rather radical hypothesis. However, from the viewpoint of observational astronomy, the concept of active galactic nuclei proved to be essential and fruitful; Byurakan astronomers have made a great contribution to the investigation of this phenomenon.

But the idea of recession within clusters is challenged by an obvious contradiction. As was shown by I. D. Karachentsev, the observed velocities of galaxies in clusters suggest that the clusters could only remain within their present-day volumes for no more than 1000-2000 million years, and therefore the clusters themselves should have existed for no longer than this period. However, the age of galaxies in the clusters is no less than 10,000-12,000 million years, which is naturally impossible if the galaxies receded within the clusters, and the galaxies were born in an explosion that produced the cluster itself.

The situation was considerably elucidated when a weighty argument that the clusters of galaxies were gravitationally related and steady-state was offered by X-ray astronomy, one of the youngest fields in astronomy, which studies celestial bodies by their emission of X-rays. In 1972, American astronomers used the specialized X-ray research *Uhuru* satellite (Small Astronomy Satellite A) to discover hot gas in clusters of galaxies. The gas temperature in the Coma cluster turned out to be very high, reaching several hundred million kelvins (this is why this gas emits primarily X-rays rather than visible light or radio waves). The thermal motions of particles at such a temperature (the gas in the clusters is ionized, and consists of hydrogen nuclei, i.e. protons) are characterized by velocities close to a thousand kilometres per second, which is close to the velocities of the galaxies in this cluster. This coincidence has been observed in both the Coma cluster and a number of other clusters of galaxies, and it is hardly accidental; most probably it implies that both galaxies and gas particles are actually "falling" in a common gravitational field with identical velocities, and this is as it should be if both the gas and galaxies are included in a single gravitationally related steady-state system.

Now many astronomers feel that they can be more confident than before in the dynamic estimate of masses. It follows that the second alternative should be accepted: we have to assume that besides luminous matter, i.e. the stars of the visible galaxies, the clusters of galaxies possess an invisible mass which controls the dynamics of these systems.

Recent data give grounds to surmise that this invisible mass, if any, is primarily to be found around major galaxies and composes their massive and extended coronas. This was deduced by J. E. Einasto and his colleagues of the Tartu observatory on the basis of the study of motions of dwarf satellite galaxies associated with massive galaxies. There are several such satellites in our Galaxy, and the Andromeda Nebula and a dozen other galaxies possess them as well. Measuring the velocity of revolution of such a satellite about the central galaxy and the radius of its orbit, we can estimate the force of gravitation between the satellite galaxy and the central mass of the galaxy.

The Tartu astronomers and then the American theorists J. Ostriker, J. Peebles, and A. Yahil noticed an amazing feature of the dynamics of satellite galaxies: the linear velocities of their revolution are the same at different distances from the central galaxy (Fig. 49).

However, it was always self-evident that these velocities should decline with the distance from the centre in inverse proportion to the square root of the orbital radius, i.e. in compliance with Kepler's law (which the solar system's planets obey). This would be exactly so if the entire mass of the central galaxy were within its visible volume as was commonly understood.

The fact that the velocities of satellite galaxies are independent of the radii of their orbits makes one think that the mass of the central galaxy is not actually limited by the stars within the bounds of its visible volume. Besides the visible stars, there should be some other gravitating masses distributed over a far greater volume throughout the gigantic galaxy. A natural guess is that these are the masses which were earlier suggested by the dynamics of clusters of galaxies.

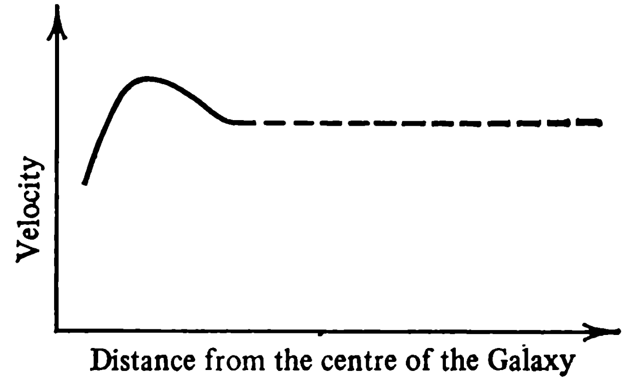

Fig. 49
A rotation curve: the velocity of the disk's rotation (solid line)· and satellite galaxies (dotted line) vs. the distance from the central gigantic spiral galaxy.

This is how the concept of invisible coronas of galaxies appeared, which primarily possess the main of the hidden masses. Coronas are spread over wide distances and reach the orbits of the farthest satellite galaxies so that these satellites move through the invisible masses rather than through a void. The gravitating mass attracting each satellite is composed of the mass of the visible stars and the hidden mass within the orbit of the given satellite.

A galaxy, its massive invisible corona, and the family of satellites define a gravitationally related steady-state system; such a system was called a hypergalaxy. A hypergalaxy obeys the general relationship between the parameters of gravitationally related systems, i.e. the virial theorem, where v, R, and M, respectively, should be understood as the velocity of a satellite galaxy, the radius of its orbit, and the mass within the orbit. If the velocity v is constant, this relationship means that the mass of the hidden matter is proportional to the distance from the centre: $M \propto R$. According to this relationship, if the mass increases, the density of matter declines in inverse proportion to the square of the distance from the centre. The density decreases very fast with the distance and finally vanishes.

During recent years, data on dwarf satellite galaxies were complemented by the data on the motion of clouds of neutral hydrogen around massive galaxies: their velocities do not depend on the radius of their orbits either. Objects of both types can be found at distances exceeding the radius of the visible body of the central galaxy by a factor of several times and sometimes a few dozen times. It immediately follows that the total mass of such a galaxy together with its corona is several times or even a few dozen times greater than the total mass of the visible stars. A typical hypergalaxy is approximately three to five times more massive than its central galaxy.

Hypergalaxies appear to be stable systems of bodies related by their common origin. Their components, from the massive central galaxy (sometimes in the centre of the system there is a sufficiently close pair of large galaxies rather than a sole galaxy) to dwarf satellite galaxies at the periphery, were formed in a single process which determined both the general makeup of the system and the structure of all its members.

Observational data on dwarf galaxies collected and studied by V. E. Karachentseva and special investigations of hypergalaxies carried out at the Tartu observatory revealed an interesting property of these systems. Satellite galaxies revolving about the centre of the system may be elliptical, spiral, or irregular; however, elliptical satellites occur far more often in the inner areas of the system than at the boundaries, while spiral and irregular satellites populate primarily the periphery of the system. A diagrammatic representation of data on the satellites of our Galaxy and three other similar galaxies, both elliptical and nonelliptical (i.e. spiral and irregular), shows that the satellites are rather clearly separated: the elliptical ones populate primarily the area to the left of the dividing line, and the nonelliptical ones to the right (Fig. 50).

This fact indicates a higher degree of regularity and organization in hypergalaxies than in common groups of galaxies. The spatial separation of elliptical and nonelliptical galaxies appears to have occurred at the formative stage of these systems when they possessed large masses of gas distributed within them. Finally, this gas settled within the central galaxy, but before this happened, it had already influenced the dwarf galaxies moving through it. If a dwarf galaxy possesses its own interstellar gas, it can retain it while moving only if the gravitational attraction exerted on the interstellar gas within this galaxy is stronger than the "counter wind" of the hypergalactic distributed gas. The "wind" may overcome this attraction in the inner area of the system, where the density of the distributed gas is greater, and therefore dwarf galaxies within this area are unable to retain the interstellar gas, and they lose it completely while making but a few revolutions around their orbits. But a galaxy devoid of the interstellar gas must look like an elliptical galaxy: there are no bright young stars in it, as is typical for spiral and irregular galaxies, which are rich in gas, but there are only old stars defining a galactic core similar to the spherical subsystem of our Galaxy. The withdrawal of the interstellar gas from satellite galaxies is less in-

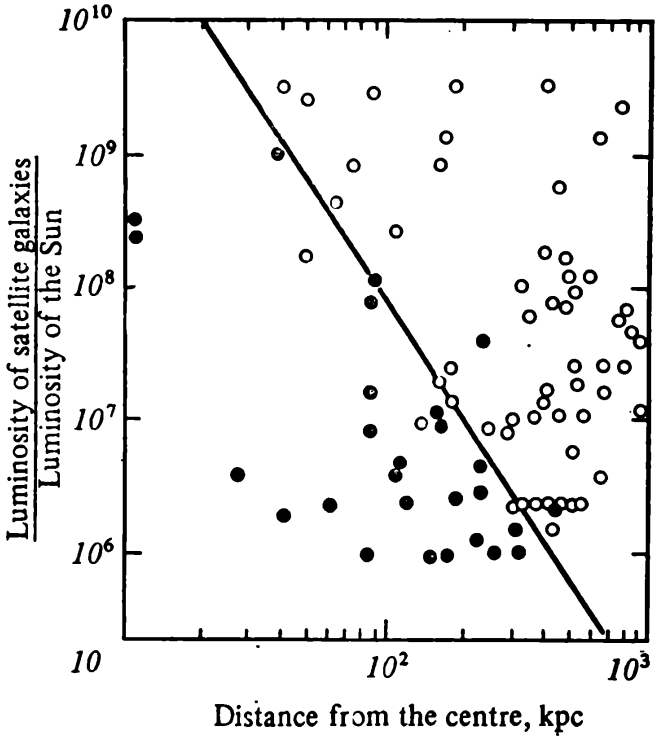

Fig. 50
Satellite galaxies on the diagram of luminosity vs. distance (according to J. E. Einasto, A. A. Kaasik, E. M. Saar, and A. D. Chernin). Logarithmic scales along both axes.

tense in the outer parts of hypergalaxies, where the density of the distributed gas is lower, and so there can exist dwarfs of both spiral and irregular types.

If this hydrodynamic process actually occurs, then there must be another peculiarity in the distribution of satellite galaxies: more massive satellites can retain the interstellar gas at shorter distances from the centre than less massive satellites because the gravitational relationship between the gas and stars is stronger in the former than in the latter. And this is precisely what can be deduced from Fig. 50; the axis of ordinates indicates luminosities of galaxies, and it is evident that they are greater for galaxies of greater masses, so more massive galaxies of all types are higher than less massive galaxies in the diagram. Spiral and irregular galaxies occur most often closer to the centre in the upper part of the diagram rather than in its lower part, which complies with what has been said above.

Hypergalaxies and clusters of galaxies are gravitationally related and steady-state because of the presence of the "hidden mass" in them. But what is the nature of this mass: cosmic dust, cool gas condensations, extinct stars, planets or perhaps black holes (the latter are always suspected whenever astronomers are challenged with a puzzling problem)...? Quite a few suggestions were made, but thorough analysis of observational data drastically reduced the number of theoretically admissible alternatives, and the following opinion gradually developed and became widespread: the hidden mass is most probably composed of very small and least bright stars that are intrinsically too faint to be observed. But the concept involving these hypothetical stars, or black dwarfs, as they are called in the literature, was not at all clear. For instance, the theory of star formation proves that such stars could not have been born within common complexes including usual stars, while mass formation of only "small" stars grossly disagrees with the accepted ideas of the cosmogonical process within galaxies.

Since 1980, everybody's attention has been drawn to quite another alternative which was proposed several years before by D. Marx and S. Salai, astrophysicists of Budapest University. They offered a hypothesis of neutrinos as the carriers of the hidden mass. For this, however, neutrinos should be particles possessing a rest mass. In the spring of 1980, a group of Moscow experimental physicists headed by V. A. Lyubimov reported the discovery of the neutrino rest mass, and the Marx-Salai hypothesis found a physical substantiation.

The Neutrino Rest Mass

Physical experiments carried out during the 1930's with radioactive nuclei established that a nucleus of tritium, a heavy hydrogen isotope, can decay and transform into a helium nucleus. While decaying, it ejects an electron and, furthermore, loses a quantity of energy. W. Pauli suggested that this energy was carried away by some unknown particles which could not be registered in these experiments (Fig. 51). They should not have any electric charge, and therefore they were similar to neutrons discovered a short time before, but these new particles were much lighter than neutrons. E. Fermi proposed to call these particles neutrinos, meaning small neutrons in Italian.

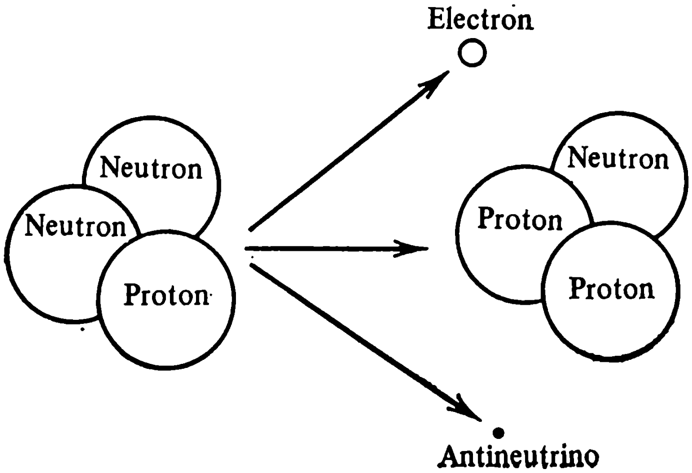

Fig. 51
The radioactive decay of tritium.

Further investigations of neutrinos revealed many of their properties, but the main point proved to be that neutrinos interact very poorly both with each other and with other elementary particles. There are four fundamental interactions in nature: electromagnetic, strong (nuclear), weak (responsible for radioactive decay), and gravitational, and neutrinos only participate in weak and gravitational interactions. In the early 1950's, physicists succeeded in registering neutrinos during direct experiments at powerful nuclear reactors, and during the 1970's, neutrinos emitted by the Sun were discovered. It was also found out that there are three kinds of neutrinos: the electron neutrino (which appears in tritium decay), the muon neutrino (which commonly comes into being together with muons, or mu mesons), and the tau neutrino (which is generated together with tau mesons).

However, it has been still not clear until recently whether the neutrino has any rest mass, i.e. the mass a particle has at rest. According to the theory of relativity, the mass of every body depends on its velocity, and this dependence becomes much more noticeable when the velocity of the body approaches the maximum possible velocity, i.e. the velocity of light. The photon, i.e. the light quantum, cannot be at rest and only exists in motion, so its mass can only be calculated from its motion. Neutrinos were in motion in every experiment that was done with them, and their rest mass could not be measured; it was only possible to establish that the neutrino rest mass is many times less than the electron rest mass, and it was thought that it might be zero like the photon rest mass.

Improved sensitivity of physical instruments allowed V. A. Lyubimov and his colleagues to find reliable evidence that the neutrino rest mass is not zero. They performed tritium decay experiments, which are traditional for neutrino physics, and neutrinos behaved as particles with a rest mass approximately 30,000 times less than that of the electron, which had until then been considered the lightest particle with a rest mass. The discoverers believe that their result is only preliminary, but if it is corroborated, it may force scientists to review certain concepts in the theory of elementary particles; the related consequences are very important for astrophysics as well.

Neutrino Coronas

According to modern cosmological concepts, neutrinos are among the most widespread particles in the universe.

On the average, there are about 450 neutrinos per each cubic centimetre of space. There are slightly fewer neutrinos than photons, but there are 1000 million times more neutrinos than protons and electrons. It is significant that this cosmological deduction does not depend on whether neutrinos have a rest mass; it follows from the primary principles of physics, and therefore there is every reason to consider it quite reliable.

Most cosmic neutrinos (and photons) are of a cosmological nature: they have not been emitted by stars or other bodies but originated together with protons, electrons, neutrons, and other particles about 15,000-18,000 million years ago. If neutrinos have a rest mass, they cannot be homogeneously scattered throughout the entire space of the universe but should cluster together, owing to their mutual gravitational attraction, into formations of certain characteristic lengths, like all other particles with a rest mass, i.e. protons, electrons, etc. They cannot condense into a planet or star, but the interstellar space of galaxies can contain them, and they are capable of clustering into a vast cloud, a corona, around a galaxy. If the estimate of the neutrino rest mass indicated above is valid, then there are enough cosmic neutrinos to fill galactic coronas and make them as massive as they appear in terms of the dynamics of galaxies and clusters of galaxies. If this is so, galaxies and their clusters are only a light luminous pattern decorating enormous formations consisting almost completely of neutrinos.

The discovery of the neutrino rest mass sheds new light on the problem of the formation of cosmic bodies from the uniform matter of the early universe. If their rest mass is as reported, then, because they are so numerous, the total mass of all the neutrinos proves to be about ten times that of all the other particles in the universe taken together. They control the general gravitational field of the cosmic medium, their own mutual attraction makes neutrinos collect into condensations of enormous mass and size, and all the rest particles (except photons) follow the course of neutrinos because they are pulled in by the gravitational field of the neutrino condensations. Consequently, neutrinos play a very important role in the cosmogony of galaxies and their clusters.

According to the theory developed by a group headed by Ya. B. Zeldovich, the succession of events during the epoch of the formation of galaxies is such that neutrino condensations appear first, and their mass is that of major clusters or superclusters of galaxies. The gas of "ordinary" particles, which is captured by these condensations, undergoes compression and heating, and then its cooling occurs, followed by the fragmentation of denser layers into protoclusters and protogalaxies. The superclusters containing many galaxies and clusters of galaxies do not disintegrate as systems but continue to exist even though they may not be as gravitationally related and steady-state as galaxies and clusters of galaxies themselves. During the initial fragmentation of the medium, neutrinos also cluster, forming what astronomical observations mentioned above reveal as the coronas of massive galaxies.

The details of this theory are a long way from being elaborated equally well; there arise rather complicated nonlinear problems of gravitational interaction between the neutrino component of the protoclusters and the gas, considering that reciprocally penetrating flows are possible within the neutrino component itself, etc. G. S. Bisnovaty-Kogan, V. N. Lukash, and I. D. Novikov have investigated most comprehensively the initial stage of the process, when the original inhomogeneities, encompassing masses of superclusters, were only weak perturbations intensified by gravitational instability.

In compliance with the data on relict radiation and the neutrino rest mass, we should distinguish three different epochs in the evolution of the primary pregalactic perturbations. During the earliest epoch, lasting no longer than a few fractions of a second after the beginning of the cosmological expansion, neutrinos collided both with each other and with other particles and exchanged their energy and pulse thanks to the weak interaction (i.e. one of the four fundamental interactions occurred, which was effective enough under the circumstances because of the great energy of colliding particles and the high frequency of these collisions). All the particles composed a single medium in the state of thermodynamic equilibrium, and the temperature of the medium was high enough for the thermal velocities of every particle to be close to the velocity of light. During the second epoch, lasting about 100,000 years, the weak interaction was insignificant because of a decline in the density and temperature of the medium during its general expansion; neutrinos did not collide and interact either with each other or with other particles any more, but they were still relativistic in the sense that the velocities of their thermal motions were comparable to the velocity of light. During this epoch, the cosmic medium consisted of two components interacting only gravitationally: collisionless neutrinos and matter mixed with light quanta. During the third epoch, which continues to date, the collisionless neutrino component has become nonrelativistic.

A significant conceptual result of Lifshits' theory is that even during the first epoch the initial perturbations were not structureless because their amplitudes (however small they might have been) were many orders greater than the level of statistical fluctuations in the medium. The nature of these triggering perturbations remains yet unknown, and only recently have physicists begun to hope to associate them with the quantum-gravitational processes in the vicinity of the cosmological singularity. By contrast, the further fate of weak perturbations during the second and third epochs can be followed confidently enough.

The development of gravitational instability requires that the gravitational attraction drawing particles closer together should not be hindered by chaotic thermal motions of particles. Gravitation is the dominating factor if the size of the perturbed area exceeds a critical value which is said to be the Jeans length.

During the first two epochs in the evolution of perturbations, the Jeans length approached the distance to the horizon. But while passing to the third epoch, when the thermal velocities of neutrinos became nonrelativistic, the Jeans length declined: see Fig. 52 showing the time dependence of the Jeans mass, i.e. the mass of neutrinos within an area of the Jeans size. The Jeans mass reached its maximum in the period between the second and the third epoch and amounted then to 4×10^{15} Sun masses.

D. Marx and S. Salai noted that this mass should have played a key role in any scenario of the cosmogonical process. In fact, it indicates the smallest size and mass of a perturbation which can grow and intensify freely at all times. This is the mass of the primary separating neutrino condensations in the theory developed by Ya. B. Zeldovich. The amplitude of perturbations for the indicated mass is considered to be the greatest during the third epoch, when a transition to the stage of strong perturbations took place, thus initiating the separation into individual masses which later gave rise to stellar systems. It is a very important fact that this mass is in effect close to the masses of clusters and superclusters of galaxies.

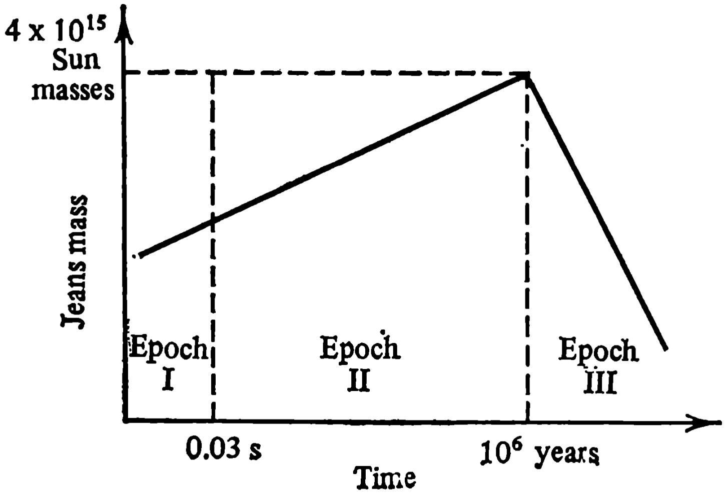

Fig. 52
The Jeans mass in the expanding universe with neutrinos possessing a rest mass Logarithmic scales along both axes.

The final stage of the cosmogonical process appears to have begun when the world was 1000-3000 million years old. One of its remarkable features is that neutrino condensations never took shape in isolation from each other but always in interaction, as if sharing between themselves all matter and condensing it into layers. These layers in turn joined and crossed each other, forming cells of irregular shapes, and the entire pattern became a quasi-ordered network structure slightly resembling a honeycomb (cf. Fig. 15, where the result of a computer simulation of such a process is presented).

During recent years, certain observational data have appeared which apparently indicate that such a cell "superstructure" does exist in the universe and is very likely of a general nature, i.e. occurring throughout the universe. A large-scale map of the universe shows that groups and clusters of galaxies are, in a number of cases, located primarily in chains or fairly thin layers that can be regarded as the walls of cells with lengths of about 100 Mpc.

The most important results of the theory of galactic formation, which were obtained earlier than the discovery of the neutrino rest mass, are completely valid in the new picture of the universe. The flattened protocluster. condensations, i.e. the "pancakes" discussed in Chapter 2, are present in this picture as well. Although the force of gravitation controlling their formation is created now not only by the gas but mainly by neutrinos, the physics of the "pancakes" does not change very much. The protocluster condensations are effectively controlled as well by the same hydrodynamic mechanisms creating vorticity in the medium and shaping the rotating gas clusters which are able to transform into spiral galaxies.

A Closed Universe?

The density of galactic matter averaged over the entire space of the universe (i.e. actually the density of baryons, or protons and neutrons) ranges approximately between 10^{-30} and 10^{-31} g/cm^3. But if neutrinos (and antineutrinos) of all three kinds possess the rest mass experimentally estimated for the electron neutrino, then their average density amounts to $(1\text{-}3)\times10^{-29}$ g/cm^3. Present-day estimates of the critical cosmological density give the value of $(1\text{-}0.5)\times10^{-29}$ g/cm^3. Therefore the neutrino density is close to the critical density. It is even possible that the neutrino density exceeds the critical density. Therefore the consideration of neutrinos possessing a nonzero rest mass can radically change our understanding of both the geometry of the universe and the fate of the cosmological expansion.

If neutrinos are not taken into account (as was always the case before the discovery by the Moscow experimenters), then the average density of the universe is definitely less than the critical density, and this suggests, according to Friedmann's theory, that the cosmological expansion is going to continue indefinitely and the volume of the universe is infinite. But if neutrinos are taken into account, the density of the universe is greater than the critical density, and the conclusion should be reversed: the expansion will not continue indefinitely and sooner or later should give way to contraction, and at the same time the volume of the universe should then be regarded as finite. Therefore a closed model of the universe should be accepted rather than the open model.

The concept that the universe is closed and that its volume is finite does not contradict the general principles of physics, but principles alone are insufficient to determine whether the universe is finite or infinite: observational astronomical criteria are required, and only they are capable of answering this question.

One such criterion is the natural requirement that the age of the universe calculated on the basis of a cosmological model should be no less than the age of the oldest stars in our Galaxy. The stars of the spherical subsystem have existed for at least 12,000 million years, and this gives the lower permissible limit to the age of the universe, i.e. it cannot be younger than this. A stricter limit, 15,000-18,000 million years, follows from the age of atomic nuclei estimated by the incidence of some radioactive isotopes. (The interested reader can find more data on the isotope techniques for estimating the age of nuclei in the book by Ya. M. Kramarovsky and V. P. Chechev (1978) listed in Recommended Literature.) It is clear that the universe cannot possibly be younger than the matter of which it consists. But this condition is not met by the closed cosmological model: the age of the universe calculated on its basis is too small, less than 10,000 million years.

It is interesting that in 1966 S. S. Gershtein and Ya. B. Zeldovich obtained a restriction on the neutrino rest mass proceeding from the requirement that the neutrino-related density of the universe has to allow the universe to have existed this long. According to this cosmological estimate, the upper limit of the neutrino rest mass was stricter than the laboratory, experimental limitations that had existed then. It was one of the first instances illustrating the profound, intrinsic relationship between the physics of elementary particles and the universe as a whole.

To date, since the neutrino rest mass has been discovered, the problem can be reversed, and the known values of mass and concentration of neutrinos can be used to calculate the age of the universe. If this calculation is carried out on the basis of the closed model of the universe, then a serious contradiction arises: the age of the universe proves, as has been mentioned earlier, to be less than the age of stars and radioactive nuclei. Moreover, if the density in the model is taken to be critical (which is also probable, as it follows from the indicated values of densities), the age of the universe is less than that of atomic nuclei.

Einstein's Vacuum

Physicists are currently attempting to overcome this contradiction and produce a new cosmological picture. They suggest advanced hypotheses on the nature of the cosmological expansion while reviewing and generalizing Friedmann's theory. A most interesting generalization of this kind is based on the concept of an unusual vacuum-like medium filling the entire universe.

A. Einstein suggested such a medium when in 1917 he applied his general theory of relativity to cosmology. He used the hypothesis of ideal regularity, i.e. that the universe has the greatest possible symmetry as a whole, as the prerequisite for his physical cosmology. This symmetry encompasses both the spatial properties of the universe and its behaviour in time. The time symmetry is the equality of all instances in the history of the universe, so that the universe is invariable and permanent. The maximum spatial symmetry is the equality of all points (the uniformity) and the equality of all directions (the isotropy) of space.

These considerations did not necessarily follow from the theory of relativity itself or from some "primary" principles or known facts of astronomy. They only generalized intuitive concepts of the global properties of the universe dating back to the science of the Renaissance: the Earth is not the centre of the universe; the Sun is one of the numerous stars scattered in space; planets and stars appeared from matter which was uniformly distributed throughout the universe beforehand (as Newton thought); any current moment in the history of the universe is an instant between the infinite past and the infinite future, etc.

The spatial symmetry of the universe is actually maximum, and to date, more than sixty years after the origination of modern cosmology, we possess reliable empirical evidence of the isotropy of physical space: relict radiation and its isotropy.

As we know to date, there is no time symmetry of the universe; the universe expands and evolves in different ways. Theory suggests that there can be no complete rest in the universe. This follows from the law of gravitation because every two particles of matter in the universe attract each other, and nothing compensates for the force of their mutual attraction, so these particles should be in motion. If the distribution of matter is generally uniform, this motion implies either general contraction or general expansion of the universe. This conclusion was obtained by A. A. Friedmann on the basis of the general theory of relativity, which is itself a generalization of Newton's law of gravitation. A. Einstein accepted Friedmann's theory completely and adhered to Friedmann's viewpoint, although he had his doubts at first.

Einstein's first response to Friedmann's paper published in 1922 in the leading international journal of physics (*Zeitschrift für Physik*) was a short critical note in the same journal indicating that Friedmann's basic deduction was erroneous. However, very soon a second note was published in which Einstein wrote, "...my criticism, as I see from Mr. Friedmann's letter communicated to me by Mr. Krutkov, was based on an error in calculations. I consider Friedmann's results correct and shedding new light."

E. Hubble's discovery of the general cosmological expansion finally corroborated Friedmann's theory and gave it a reliable observational foundation. Later, summarizing the theoretical investigations on which modern cosmology rests, Einstein wrote that "Friedmann was the first to pave the way."

Experiments and observations are crucial in theoretical discussions. Thus, the theory of the expanding universe proved to be valid, while the ideas of the static and permanent universe were abandoned, although the Einstein theoretical model of the static universe did not contain any direct error (the above-quoted note mentioned an error in calculations, but this referred to verifying calculations carried out while analyzing Friedmann's paper rather than to Einstein's own theory).

The Einstein model contained a very important ele-. ment, a bold hypothesis making it possible to combine the static universe with the law of gravitation. The static nature of the universe required that besides the gravitating matter, there should be an extraneous force factor capable of compensating for the forces of mutual attraction between all bodies within the entire universe taken as a whole. Naturally, everything was "as before" in the solar system, in individual stars, and galaxies, and there was no compensation for the gravitation of these bodies. It was the field of the gravitation of the cosmological scale produced by the generally uniform (on the average) distribution of matter that should be compensated for. This factor was defined as a hypothetical vacuum, a uniform medium characterized only by antigravitation, i.e. the ability of the bodies within it to repel each other.

This medium possesses mass and energy, but nevertheless the antigravitating medium described by Einstein is referred to as a vacuum because there are no real particles in it. Furthermore, it possesses a special property: there is no distinction between motion and rest with respect to this medium. For instance, we know that while a body moves with respect to a gas, there appears a "counter wind" of particles, atoms or molecules, and we can count the number of particles carried by this wind per unit time through a unit cross-section, and find the velocity of the wind and therefore the velocity of our motion with respect to the gas. There is no "counter wind" in the motion through Einstein's medium; so the medium is at rest with respect to us, and we are at rest with respect to the medium, no matter what the velocity and direction of our motion in space might be. This is a very special property of the medium, but this is precisely a property of a vacuum: there is no "wind" in a void. Consequently, Einstein's medium possesses the mechanical properties of a vacuum, i.e. the properties that reveal themselves with respect to the motion of bodies (this was pointed out by E. B. Gliner).

This property of Einstein's vacuum is described by a certain relationship between its mass density ρ_v, or energy density $\varepsilon_v = \rho_v c^2$, and the pressure p_v which has also to be ascribed to this medium: $\varepsilon_v = -p_v$. However, no "normal" medium has the pressure whose modulus is equal to its energy density and is opposite to it in its sign. But the medium does not produce any "counter wind", no matter what our displacement in it might be, if and only if this relationship exists between its pressure and density.

Another, no less amazing property of the antigravitating medium (linked to the relationship stated above between its pressure and density) is its complete uniformity in space and permanence in time. The pressure and density of Einstein's vacuum are everywhere identical and do not vary with time. This was the property of the antigravitating medium in Einstein's static, ideally symmetrical cosmological picture. The antigravitating medium retains this property in the theory of the expanding universe: the density and pressure of "normal" matter decline with the expansion, but the vacuum remains the same.

The vacuum density does not vary with time, and in fact, this implies that we are dealing with a new fundamental constant. Its value should be determined by means of astronomical observations, and if it actually proves to be nonzero, the problem of the existence of this unusual omnipresent medium will be solved. These observations should be carried out in the future. Today we can only suggest that most likely the density ρ_v cannot greatly exceed the present-day density of matter in the universe. A plausible model we are going to discuss below holds that $\rho_v \lesssim 10^{-28}$ g/cm^3. (Recall that the neutrino-related density amounts to about 10^{-29} g/cm^3.)

However, it should be noted that the indicated value is basically the upper limit of the vacuum density: a vacuum cannot have a density exceeding the indicated value, but this does not prevent it from possibly being zero.

Einstein's vacuum is permanent and invariable, and it has always been perplexing. It represents a medium which influences other media and creates forces controlling the motions of bodies, whereas the medium itself is not subject to any influences or counteractions! This is definitely strange, and there has so far been no explanation for it. Possibly, it will become clear when the nature of Einstein's vacuum is established and then perceived and comprehended on the basis of general concepts of the physics of elementary particles.

As we have already mentioned, the concept of the antigravitating vacuum is compatible with the expanding universe. Therefore it is possible to generalize the "standard" Friedmann cosmological models. A. A. Friedmann himself suggested them in his classical work during the period from 1922 to 1924. G. Lemaître analyzed these models in detail in the 1930's, and he contributed significantly to the study of the relationship between cosmological models and real astronomical observations.

Returning to the problems arising in cosmology in connection with the discovery of the neutrino rest mass, let's turn our attention to the fact that models with an antigravitating vacuum have different relationships between the density of the universe and its age. A new "free parameter", the vacuum density, appears in them instead of the unambiguous relationship between these two values in the "conventional" Friedmann models. In principle, this parameter can be chosen so that the age of the universe is sufficiently great. Figure 53 illustrates this possibility: in a generalized model of a certain type (called Lemaître's model), a given density (the horizontal broken line) can correspond to an age of the universe which is noticeably greater than that of the Friedmann model without a vacuum. The density of Einstein's vacuum in this case is comparable to the density of matter, i.e. actually that of neutrinos: $\rho_v \simeq (10^{-29}\text{-}10^{-28})$ g/cm^3.

It is interesting that the relationship between the geometry of the universe and its dynamics in cosmological models with Einstein's vacuum is more ambiguous than in the "conventional" Friedmann model, in which the open universe corresponds to infinite expansion in time, whereas the closed universe implies that the expansion will eventually give way to contraction. The volume of the universe is finite in Lemaître's model, the universe is closed, but its expansion is always infinite in time. This model of the universe seems to agree with all astronomical data and experimental physical facts.

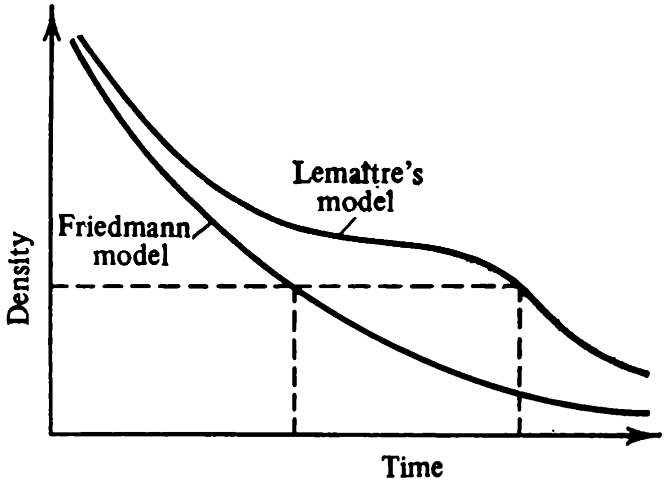

Fig. 53
The density of the universe vs. its age in the "standard" Friedmann model and in Lemaître's model.

However, this is not the ultimate answer to the question of whether the universe is finite or infinite. This question has challenged the human mind from time immemorial. Here is an English rendering of what M. V. Lomonosov, a Russian encyclopaedic natural scientist, wrote in 1743:

> Your answer's full of subterfuge
> About what's not near but far
> Can the universe be so huge?
> And what's beyond the faintest star?

Conclusion

Writing this book, we wanted to give the reader the most important and interesting facts and ideas concerning a new field of astrophysics, the cosmogony of galaxies and stars. Galaxies have been subjects of cosmogonical investigations since the 1920's, when their true nature was reliably established and they proved to be vast stellar worlds beyond our stellar system at very great distances from the Milky Way (our Galaxy) rather than nebulae, i.e. small clouds of gas and dust nearby. The names of some galaxies have retained until now traces of once widespread views and concepts. For instance, the closest gigantic spiral galaxy is still called the Andromeda Nebula. The elaboration of a sequential picture of the appearance and evolution of stars was also made possible only during the past fifty years owing to the successes of observational astronomy and the development of the physical theory revealing the nature of stellar luminosity. The discoveries and investigations in the field of cosmology during the past few decades have elucidated many points concerning the prehistory of galaxies and stars and the physical state of the rarefied matter out of which they formed very long ago.

The reader sees that the basis of modern cosmology is the fundamental idea dating back to I. Newton: the idea of gravitational instability. Matter cannot remain uniformly diffused in space because the mutual attraction between all particles of matter tends to produce condensations of certain scales and masses. The gravitational instability in the early universe intensified the initially very weak irregularities in the distribution and motion of matter, and during a certain epoch generated the appearance of strong inhomogeneities: protocluster "pancakes". The boundaries of these condensed layers were shock waves at whose fronts the originally nonrotatory, eddy-free motion of matter acquired vorticity. The separation of layers into individual condensations also appears to have occurred because of gravitational instability, and this gave rise to protogalaxies. Many of them possessed fast rotation because of the eddies in the matter out of which they are formed. The fragmentation of protogalactic clouds due to gravitational instability led to the appearance of the first stars, while clouds turned into stellar systems such as galaxies. The galaxies possessing fast rotation acquired a two-component structure consisting of a halo of more or less spherical shape and a disk with spiral arms where stars are born even now. The protogalaxies whose rotation was slow or completely absent turned into elliptical or irregular galaxies. Concurrently with this process, the formation of the large-scale structure of the universe was under way: superclusters of galaxies appeared, which jointly form something like cells or a honeycomb. These cells have only been discovered during recent years.

This is a general outline of the cosmogonical process in the universe as it is dealt with in this book. The only thing left to do is to tell the story of the appearance of a common star, one of the few thousand million stars in our Galaxy, but the most important one for us. So, in conclusion we shall give an account of the appearance of the Sun and the solar system.

During recent years, the solar system has become the subject of direct experimental rather than purely observational investigations. Unmanned space missions, orbital laboratories, and manned missions to the Moon yielded numerous new concrete data on the Earth, nearby space, planets, and the Sun. Consequently, the development of the cosmogony of the solar system could have an entirely new basis.

The origin of the solar system was historically the first problem of cosmogony formulated as a problem of natural sciences. In 1644, R. Descartes put forward the concept of a protosun nebula. It was a rotating cloud of eddies of gas and dust. The Sun formed at the centre of the cloud, and the planets with their natural satellites appeared at the periphery. A century later, I. Kant and then P. Laplace, developing this idea, investigated the dynamics of the rotating cloud on the basis of Newtonian mechanics. Compressed by its own gravitation, the cloud rotated faster and faster and flattened into a disk; at a certain stage, fast rotating rings successively separated from the rim of the disk owing to inertia. Then the material of individual rings condensed and produced planets, each in its own orbit. Thus appeared a visual explanation of the most important property of the solar system, i.e. that its planets move around almost circular orbits, the orbits are in the same plane, and the planets revolve around their orbits in the same direction which is identical to the rotation of the Sun about its axis.

It is far more difficult to explain other regularities of the solar system, primarily the distribution of the angular momentum between the Sun and the planets. The planets possess about 1/700th fraction of the Sun mass. However, the angular momentum related to their optical motion amounts to 98 per cent of the entire angular momentum of the solar system. (The angular momentum associated with the proper rotation of the planets about their axes is insignificantly small compared to the momentum of their orbital motion.) The proper motion of the Sun about its axis yields only 2 per cent of the total angular momentum of the system. In other words, the rotation for some reason proved to be very unevenly distributed with respect to the mass of the system: the greatest mass is in the centre, while most of the angular momentum belongs to the periphery.

It is noteworthy that Jupiter possesses an angular momentum exceeding that of the Sun by a factor of over thirty times. (The estimate of Jupiter's angular momentum is its mass 2×10^{30} g, times the radius of its orbit 7.8×10^{13} cm, times its orbital velocity 1.3×10^{6} cm/s.) Saturn's angular momentum is less than half that of Jupiter's, while the momenta of all the other planets taken together amount to less than a quarter of Jupiter's angular momentum.

This distribution of the angular momentum cannot occur in a successive separation of equatorial rings from a rotating cloud. The problem of momentum also remained unresolved in a number of other cosmogonical models proposed in the past and still discussed today. In 1745, J. Buffon suggested that the matter of the planets was drawn from the Sun by a celestial body travelling nearby, e.g. by a comet. J. Jeans and other cosmogonists following in his footsteps, from the 1920's to the 1940's, thought that the matter of the planets could be drawn from the Sun (which could have been formed by that time) because of a close approach of a neighbouring star. It was suggested that the gravitation of the approaching star created a jet of matter flowing from the Sun. The matter of the jet was nevertheless gravitationally bound to the Sun, and when the star withdrew, the planets formed out of the material of the jet.

However, a close encounter of stars accompanied by a flow of matter is an event of very low probability. We do not have any data on other planetary systems because we cannot see any other system besides ours, so there is possibility of the uniqueness of an event of this kind. But the assumption of a very close encounter of stars seems to be very far-fetched from the viewpoint of theory. It is also essential that the conditions for the condensation of the planets in the jet drawn from the Sun appear to have been unfavourable for its matter to gather into protoplanets rather than scatter.

During the 1940's, O. Yu. Schmidt and then the Swedish physicist H. Alfven investigated the possibility of capturing protoplanet material while the Sun passed through gas-dust clouds in the disk of the Galaxy. It appears that the problems of the condensation and the angular momentum are the same in this picture as well.

A wide spectrum of ideas and opinions have been collected in the book *The Origin of the Solar System* (1972) edited by J. Reeves. The initial point of many modern cosmogonical schemes remains to be the concept of a single primordial rotating gas-dust cloud from which the Sun and the planets formed. This concept dates back to the classical Descartes-Kant-Laplace hypothesis, and most likely it can help solve such problems of cosmogony as the chemical composition of the Sun and the planets, as well as the problem of the condensation of matter in the protoplanet cloud. New ideas give hope of solving the key problem, that of the angular momentum.

The starting point of modern studies of the cosmogony of the solar system is that the material of the protoplanet cloud was the interstellar medium, whose chemical composition differed little from the composition of the present-day interstellar medium. Heavy elements, whose mass amounts to about 2 per cent, were contained primarily in the dust particles; the medium was 70-75 per cent hydrogen and about 23-28 per cent helium. While the cloud was compressed by the force of its own gravitation, its rotation intensified, and eventually inertia impeded the contraction at right angles to the axis of rotation. However, the contraction along the axis of rotation continued, and the dust particles, i.e. small solid bodies, settled in the median, equatorial plane of the cloud faster than the hydrogen-helium gas did. The dust particles accumulated in the median plane of the cloud, and therefore collisions between them should have occurred there rather frequently, so the particles could have coalesced into larger solid bodies. A rather flat disk appeared consisting mainly of such bodies and separate dust particles.

It appears that along with the formation of the disk another process took place: a dense gaseous globe shaped in the centre of the cloud, then the globe turned into a star, and thermonuclear reactions started within it. This star, the Sun, began to heat the surrounding matter, causing hydrogen and helium to escape gradually from the region close to the Sun to the outer regions of the cloud. The chemical composition of the protoplanet cloud became inhomogeneous: it divided into a region close to the Sun, where heavy elements prevailed (they were collected into dust particles and then solid bodies were produced out of them), and the periphery, where the chemical composition remained almost as it had been before.

This is why the planets, which were formed out of the material of the protoplanet cloud, are so different in their physical properties and chemical composition. Gaseous globes, i.e. the planets consisting mainly of hydrogen and helium, including Jupiter, Saturn, Uranus, and Neptune, appeared in the outer region, at the periphery of the solar system. Solid planets, where elements heavier than hydrogen and helium prevailed, appeared in the inner region. Volatile substances, primarily hydrogen and helium, escaped the region within the orbit of the

Earth almost completely, and therefore the Earth, as well as Mars, Venus, and Mercury, are solid.

L. E. Gurevich, A. I. Lebedinsky, B. Yu. Levin, and V. S. Safronov developed a detailed picture of the fragmentation of the protoplanet cloud into different layers. It was found that small condensations (planetesimals) rather than planets appeared in the process. Gradually joining each other, these planetesimals collected into planets, and this was the final stage of the formation of the solar system.

However, the problem of the observed distribution of the angular momentum in the solar system persists. An interesting idea has been recently suggested by E. M. Drobyshevsky. His hypothesis proceeded from a rather unexpected result of digital computer simulation of the evolution of a rotating cloud under the effect of its own gravitation. This numerical investigation was conducted by R. Larson, and it showed that while the cloud contracts and its rotation increases, it transforms into a rotating torus rather than a smooth flat disk. But a rotating torus without a central mass is a totally unstable configuration (this was proved by V. A. Antonov). It should disintegrate into separate condensations revolving around their common centre of mass.

E. M. Drobyshevsky suggests that most likely there were two comparatively large condensations and many other, less massive ones. One of the major condensations became the Sun, and the other became Jupiter, the most massive planet of the solar system. But before this happened, there had occurred a complex interaction between the two major condensations with a flow of matter from one of them to the other (for instance, as in the close binary system of a burster). Matter flowed as a jet from the future Jupiter to the future Sun. An accretion disk formed around the protosun, whose rotation could be opposite to the axial rotation of the protosun. When the material of the disk settled on the surface of the protosun, it slowed down its axial rotation. This is the reason why the rotation of the Sun is so slow, and its angular momentum is respectively small.

However, quantitative estimates are still not quite definite, and apparently it will take great efforts before the problem of the angular momentum of the solar system can be resolved.

Astronomical observations in the search for other planetary systems are going to produce more data. Until now, there has been no observational answer to the important question of whether the solar system is a common or an exceptional phenomenon. It is only known that one of the stars closest to the Sun, Barnard's star, possesses an invisible satellite.

This satellite cannot be a star (its mass is too small for that) and therefore is a planet. Small satellites of more distant stars cannot be discovered by the state-of-the-art equipment. Barnard's star is at a distance of 1.83 pc from us, and the fact that the only close star, for which such a discovery was possible, possesses a planet-like satellite makes likely the assumption that they are a common phenomenon for stars of certain types.

But back to the solar system. We shall conclude with the story of our planet, i.e. the Earth. The planetesimals out of which it was shaped collided and therefore coalesced, became heated, and therefore melted. Possibly, this is how the hot molten core of the Earth originated. The present-day Earth also has a molten core, which is heated now owing to the radioactive decay of heavy atomic nuclei. The Earth's core consists primarily of iron (with an admixture of nickel), while hydrogen, silicon, magnesium, and their compounds surfaced from the hot melt long ago, cooled, and formed the solid mantle surrounding the core. Siliceous rocks of lesser density produced continents, while the lighter elements and their compounds gave rise to the oceans and the atmosphere.

The atmosphere and the oceans of the Earth were the environment where life originated. The mixture of the atmospheric gases was conducive to the appearance of complex organic molecules. This process was enhanced by the ultraviolet radiation of the Sun and possibly by the violent thunderstorms frequently occurring in the atmosphere. Probably, this is how many amino acids and other organic compounds appeared. Precipitating into the ocean along with rain, they were capable of joining there into still more complicated long chains of molecules, i.e. proteins and nucleic acids, from which the first living cells developed, and then some primitive forms of vegetation, such as algae, appeared. Later, plants occupied the dry land as well.

At the same time, gradual changes occurred in the atmosphere. Its initial composition differed greatly from its present-day composition: there was a lot of hydrogen, helium, and hydrogen-containing gases: ammonia, methane, and water vapour. Hydrogen and helium, the lightest elements, could escape gradually, while oxygen appeared in the atmosphere, produced mainly by plants in the process of photosynthesis. "Exhaling" oxygen, plants created favourable conditions for the origination of animal life on our planet.

Is it possible that a similar evolutionary process, leading to the appearance of life and then to forms of higher intelligence, could take place on other planets, in other stellar systems? It appears that there are no grounds to deny such a possibility. The opinion that intelligent life exists not only on the Earth was considered heretical several centuries ago, but now it is almost universally accepted. It is held both by believers in "flying saucers" and scientists who have been thoroughly analyzing in recent years the options for a directed search for extraterrestrial civilizations. However, there are no reliable data on even the most primitive forms of life on other planets.

We do hope that the reader agrees that both of the two alternatives, i.e. either the existence of numerous inhabited worlds or the unique occurrence of intelligent life only on the Earth, are exceedingly dramatic, though provoking, and intriguing....

Contents